Bayesian methods for validating patient reported outcomes and

predicting patient accrual in clinical trials

By

Yu Jiang

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy

_____

Chairperson Byron J Gajewski, Ph.D.

_____

Marge Bott, RN, Ph.D.

_____

Francisco Diaz, Ph.D.

_____

Jianghua He, Ph.D

_____

Jo Wick, Ph.D.

Date defended: September 5, 2014

The Dissertation Committee for Yu Jiang certifies that this is the approved version of the following dissertation

Bayesian methods for validating patient reported outcomes and

predicting patient accrual in clinical trials

_____

Chairperson Byron J Gajewski, Ph.D.

Date approved: September 5, 2014

ABSTRACT

Bayesian approaches to the design and analysis of clinical trials and health-care evaluation studies are becoming popular. Compared to frequentist methods, the Bayesian approach offers a flexible way to incorporate all sources of information into an analysis, including both preexisting knowledge as well as new information and evidence as it occurs. Bayesian methods have been shown to be more efficient when using appropriate informative priors, and have been used in various clinical studies and health-care evaluation studies. In this dissertation, we have developed Bayesian models and applied these models to two specific problems: (1) the development and assessment of patient reported outcomes and (2) the prediction of patient accrual in clinical trials. Bayesian Instrument Development (BID) is introduced to examine patient reported outcomes. The stability of BID is evaluated using simulation study, and user-friendly BID software is also proposed. The BID software can provide Bayesian estimates of content and construct analysis for developing patient reported outcome measures. Also, we developed and tested Bayesian models and applied them to patient accrual monitoring in clinical trials. Specifically, two hierarchical priors are introduced, and the properties of different priors are evaluated using data from clinical studies and simulations. User-friendly software for accrual monitoring is described and details for its use are provided. The accrual software is used to predict the accrual at the end of a clinical trial given accrual to the present time, and can provide the probability that the trial will finish within the planned time frame or the time frame required to recruit the planned number of subjects. The dissertation concludes with summary and future studies.

ACKNOWLEDGMENTS

At first, I sincerely thank my major advisor, Dr. Byron J. Gajewski for the wonderful opportunity he gave me to join his research group and pursue my doctoral degree. It is his wisdom advice, excellent supervision, numerous encouragements that make me finishing my study and pursing my career possible. I appreciate his endless patience, guidance and support all over the years.

My thanks also give to my graduate committee members, Dr. Jo Wick for her great advice and wonderful instructions for my study; Dr. Francisco Diaz for his excellent guidance and valuable comments; and Dr. Marge Bott, and Dr. Jianghua (Wendy) He for their comments and suggestions all along with my study.

Specially, thanks to Dr. Matt Mayo, who admitted me to the program and supports me all the time. I also want to thank our Graduate Education Coordinator, Jackie Jorland. She is always there helping me. She is the one that makes my study in the department going more smoothly and more joyful.

I want to thank my colleagues: Lili Garrard, Yang Lei, Janelle Noel, Wei (Will) Jiang Milan Bimali, Xueyi Chen, Their friendship and their kindly help in my study were valuable and unforgettable.

At last, I would thank my dear husband, Lei Dong. It is his love and devoted support that make me to pursue my career possible. I would also love to thank my son Joshua, for the all happiness he brings to me. Thank you, all my families and friends, for all your help and kindness!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE:

INTRODUCTION

Reverend Thomas Bayes first proposed a theorem to update beliefs that later became known as his 'rule' in the 1740s, though the result was not published until two years after his death (Bayes, 1763). In the 2½ centuries since, Bayesian statistics has evolved from controversial to widely used for solving real-world problems in politics, economics, engineering, and the basic and biomedical sciences. For example, Turing applied Bayes' rule to break the German cipher Enigma and locate U-boats during World War II (McGrayne, 2011). More recently, Stone et al. (2014) used a Bayesian approach to locate the debris of Air France Flight 447 which crashed in the Atlantic Ocean on June 1st, 2009. Beginning with the rapid development of computational tools in the 1980s, Bayesian methods have become an increasingly popular and widely accepted approach to statistics. Thus, it is not surprising that healthcare research also has benefited from Bayes' rule.

Bayesian approaches to the design and analysis of clinical trials and health-care evaluation studies are becoming popular (Berry, 2004). Compared to frequentist methods, the Bayesian approach offers a flexible way to incorporate all sources of information into an analysis, including both preexisting knowledge (e.g., historical information, researchers' previous experience) as well as new information and evidence as it occurs (Spiegelhalter, Abrams & Myles, 2004). Bayesian methods have been shown to be more efficient when using appropriate informative priors (Samaniego & Reneau, 1994), and have been used in various clinical studies and health-care evaluation studies (Berry, 2004).

Bayesian methods, like frequentist methods, require a data model (or likelihood), $f\left(\mathbf{y}|\theta\right)$, where $\mathbf{y}$ represents observable quantities (e.g., experimental data) and θ denotes unobservable model parameters. Relevant prior information about θ is incorporated into the analysis by means of a prior probability distribution, represented by $\pi\left(\theta\right)$. Information about

θ is subsequently updated by the data generated from the current experiment through the likelihood, $f(\mathbf{y}|\theta)$. This updated knowledge about θ is represented by the posterior probability distribution, $\pi(\theta|\mathbf{y})$, where Bayes' theorem establishes the relationship between the prior and posterior information:

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int_{\Theta} f(\mathbf{y}|\theta)\pi(\theta)d\theta},$$

where Θ denotes the parameter space (Casella and Berger, 2002). Specification of prior distributions, model assessment, and use in practice are key elements of Bayesian statistical analysis that will be addressed in the following sections.

1. Bayesian prior distributions

In general, Bayesian prior distributions can be categorized as non-informative or informative based on the amount of information they provide about θ relative to the information provided through the data likelihood. Non-informative priors—sometimes referred to as reference or flat priors (Gelman et al., 2013)—are generally uniform over the range of values for θ that are of interest. Uniform densities, by definition, provide little information about θ aside from a range of possible values, thus allowing the data to calibrate the posterior distribution over the most likely values of θ through the likelihood. Therefore, non-informative priors provide data-driven analyses and the results obtained from such approaches are very similar to those from corresponding frequentist methods.

Unlike non-informative prior distributions, informative priors incorporate into an analysis preexisting knowledge or expert beliefs about the most likely values of θ. Eliciting expert beliefs and translating them into a prior distribution in mathematical terms is challenging and

considered to be the most important step in Bayesian statistics (Spiegelhalter, Abrams & Myles, 2004). Common approaches of prior elicitation have been summarized and discussed by many researchers (e.g., Spiegelhalter, Abrams & Myles, 2004; Johnson et al., 2010; O'Hagan et al., 2006). According to Spiegelhalter, Abrams & Myles (2004), the different approaches can be divided into: (a) informal discussion, (b) structural interviewing and formal pooling of opinion, (c) structured questionnaires, and (d) computer-based elicitation. Through these approaches, various priors have been identified and are commonly applied in Bayesian analyses (Spiegelhalter, Abrams & Myles, 2004). As prior beliefs of experts come from different expectations, informative priors can be either "pessimistic" or "optimistic" (Spiegelhalter, Abrams & Myles, 2004; Cook, Jairo, & Pericchi, 2011). Pessimistic priors reflect a skeptical position and usually are centered over values of θ with high probability of corresponding to the null hypothesis. Optimistic priors relate to more positive expectations and correspond with values of θ that reflect the hypothesis an investigator wishes to support through experimentation.

Because of the subjective nature of informative priors based on expert opinion, there have been many criticisms of the prior eliciting process. One of the criticisms is that experts could be biased in their beliefs. Studies have shown that many investigators tend to overestimate the strength of their prior beliefs; that is, expert opinion tends to be overly "optimistic" (O'Hagan et al., 2006). For example, it was shown by Hughes that clinicians tend to expect that a new therapy /drug to be a benefit in a clinic trial study. Therefore their prior brief on the new therapy usually is exaggerated (Hughes, 1991). The bias of prior elicitation also may come from the choice of experts and/or time of elicitation (Kadane & Wolfson, 1998). When priors are miscalibrated,

Bayesian estimates may lose their superiority to other estimation methods and should be processed with caution (Samaniego & Reneau, 1994).

To overcome the concerns associated with expert-based informative priors, methods have been proposed to incorporate historical information effectively and less subjectively (Berger, 2006; Ghosh, 2011). Power priors (Ibrahim & Chen, 2000) are one such method that avoids many of the difficulties and criticisms in prior elicitation by permitting prior parameters to be determined directly from historical information. Because the original form of the approach violated the likelihood principle, a modified power prior was proposed by several groups of researchers (Duan, Ye, & Smith, 2006; Neuenschwander, Branson, & Spiegelhalter, 2009):

$$\pi(\theta, a_0|D_0) = C(a_0)\, L(\theta|D_0)^{a_0}\pi_0(\theta)\pi(a_0).$$

Specifically, the modified power prior contains two parameters: historical information, D_0, and a weight parameter, a_0. The borrowing of strength from the historical information is controlled by raising the likelihood of D_0, $L(\theta|D_0)$, to a power, a_0, representing the relative importance of the historical information to the Bayesian analysis. This weight parameter is typically restricted to values between zero and one, with a_0 = 1 corresponding to the usual updating associated with Bayes' theorem. As a_0 approaches zero, the impact of the historical information on the power prior distribution diminishes.

Another way to utilize historical data objectively is through the commensurate power prior, which is proposed by Hobbs et al. (2011). The commensurate prior assumes different parameters represent historical (θ_0) and current (θ) information. The precision, τ, of θ conditioned on θ_0 parameterizes the commensurability (correspondence) between the two sources of information. For example, when evidence for the commensurability is weak, τ will be small and the conditional prior variance of θ will be increased. On the contrary, when the

current and historical data are similar, the commensurability is strong and the precision will be large.

The idea of objective priors in Bayesian statistics has been debated (Goldstein, 2006; Berger, 2006; Kadane, 2006). They are not recognized as "fully Bayesian" by some Bayesian statisticians and it is advised that the priors should be used with caution. In all cases, researchers should evaluate the appropriateness of the model (Kadane, 2006; Press, 2009). As stated by Spiegelhalter, Abrams, & Myles (2004), there is no concept of a "correct" prior distribution—the properties of different priors could be different under varying situations. Therefore, model evaluation is critical in guiding and assessing the impact of prior specifications.

2. Model Evaluation

By definition, statistics is the science of decision-making in the presence of uncertainty. Due to the inherent uncertainty in measurements taken from samples it is near impossible to obtain a perfect statistical model. In general, statistical models are known to be imperfect; therefore, it is important to compare competing models. An ideal statistical model would be simple, yet representative and useful in real situations.

Statistical models can be evaluated using real data or via simulation studies. Mean squared error (MSE) is a frequentist criteria for model evaluation that represents the squared difference between an estimator and its target that can be used to identify which of a set of candidate models best fits the data. A model or method is shown to be superior if it provides smaller mean square error. MSE incorporates both the variance and bias of an estimator. To illustrate, suppose the hits on a target (the true value of a parameter) represent the model building process to estimate a parameter, with each hit symbolizing an individual realization of the model system. Figure 1-1 displays the different realizations for four different scenarios in the form of

scatter plots. It always is ideal to find an unbiased estimator with low variance (Figure 1-1A). However, the unbiased estimator does not guarantee an efficient estimator; that is, an estimator that requires fewer samples to achieve a prespecified level of precision (Figure 1-1B). Alternatively, an estimator will be superior if it has negligible bias but smaller variance (Figure 1C). In this way, MSE incorporates both the degree of variation and the degree of bias in an estimator to assess its quality.

Evaluation of the Bayesian approach and comparisons to frequentist alternatives have been studied extensively (Chaloner, 1987; Samaniego & Reneau, 1994; Browne & Drapper, 2006; Neath & Langenfeld, 2012). It was first shown theoretically and empirically by Samaniego & Reneau (1994) that Bayesian estimates are better than a likelihood approach when the priors are calibrated closely enough to the truth. When a small bias exists, Bayesian methods still outperform frequentist methods if the data has a paradoxically large variance. When prior distributions are highly miscalibrated, Bayesian methods will no longer be superior to other methods. As previously discussed, prior specification is very important in a Bayesian approach. Therefore, it is also critical to do model comparisons between Bayesian methods for different prior specifications. An ideal prior setting would be robust and work well under varying conditions.

3. Software

It is the significant improvement of computational techniques that spurred the rapid development of Bayesian statistics and its increased use in different fields, especially in areas of the design and operations of clinical trials and health-care evaluation. With the availability of high speed computers, Bayesian analysis can be constructed, validated, and used in practice. There are several software packages and languages that can be used. For general computing,

statisticians write their own Bayesian algorithms using a lower level language, such as FORTRAN and C++ (Eubank & Kupresanin 2011). Some functional languages, such as R, MATLAB, Phyton are very popular and used a lot by statisticians (Rashed et al., 2012; Albert, 2007). WinBUGS (Lunn et al., 2000; Lunn et al., 2009), OpenBUGS (Spiegelhalter, et al., 2007), JAGS (Plummer, 2003) and STAN (Stan Development Team, 2014) are higher level functional languages specific for Bayesian analysis using Markov chain Monte Carlo (MCMC) methods and/or other optimization algorithms. SAS has some Bayesian options within its procedures but requires users with in-depth knowledge of statistical computing in SAS (Sullivan & Greenland, 2013). SAS also is not very flexible for implementing new and innovative approaches. Other well-known graphical user interface (GUI) statistical software (e.g., SPSS and JMP) do not have Bayesian capabilities. In general, Bayesian statistical software is limited, especially for clinical researchers.

For most clinical researchers, it is not practical to implement a Bayesian analysis using a functional language, such as R and WinBUGS, unless easy-to-use platforms such as GUIs can be developed. Therefore, to make the Bayesian approach more feasible for use by clinical and biomedical researchers, easy-to-use tools for conducting Bayesian analysis must be developed. Tools that require minimal interaction—such as input the experimental data and simple point-and-click—while running the Bayesian prediction model in the background would have the most potential for real-world problem solving.

4. Specific studies

In the current studies, we have developed Bayesian models and applied these models to two specific problems in health care and clinical study: (a) the development and assessment of patient reported outcomes and (b) the prediction of patient accrual (recruitment) in clinical trials.

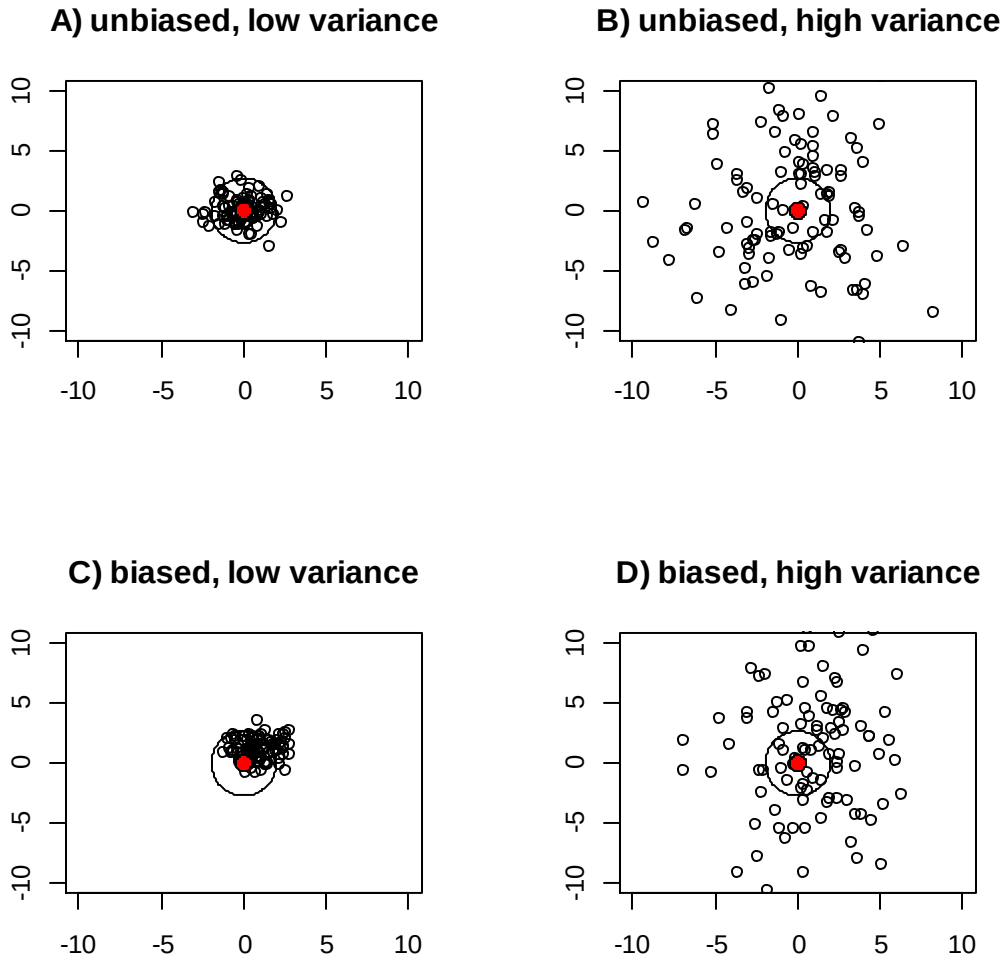
Patient-reported outcome measures are gaining in importance for the evaluation of clinical studies (NIH; Cella et al., 2010; FDA; Speight & Barendse, 2010). For some chronic diseases, the ultimate goal of clinical treatment not only is disease survival but the quality of life. In addition to physical, physiological or biochemical measures of a disease condition, patient-reported outcomes on feedback and behavior changes provide crucial information regarding the treatment of disease (Deshpande, 2012). Unlike other clinical outcomes, such as blood pressure, fasting glucose, and lipid profile, patient reported outcomes cannot be measured by medical devices. Instead, they require a psychometric instrument (e.g., a questionnaire) with supporting documents. Constructing a valid instrument is both time consuming and expensive, especially in situations where measurement development requires patients thatcome from small populations (e.g., American Indians) and/or suffer from orphan diseases (e.g., cryptogenic painful neuropathy). A valid and efficient instrument development method will help to expedite the process and thus be beneficial for studying the impact of treatment and disease on patients.

Subject recruitment is another challenge in medical research. Slow patient accrual leads to increased cost and resource utilization, especially the goodwill contribution of patient volunteers. When accrual is slow, researchers potentially will settle for smaller sample sizes than originally proposed. Consequently, slow accrual could result in a delay in the adoption of new therapies and slow the advancement of medical progress (Philipson et al., 2010). Overall, clinical trial monitoring of the accrual process is very important.

In the following chapters, Bayesian methods will be developed and applied to problems in health care evaluation. In Chapter 2, Bayesian Instrument Development (BID) is introduced to examine patient reported outcomes. The stability of BID is evaluated using a simulation study, and user-friendly BID software also is proposed. The BID software can provide Bayesian

estimates of content and construct analysis for developing patient-reported outcome measures. Chapters 3 and 4 are studies of patient accrual monitoring in clinical trials. In Chapter 3, two hierarchical priors are introduced, and the properties of different priors are evaluated using data from clinical studies and simulations. Also in Chapter 3, user-friendly software for accural monitoring is described and details for its use are provided. The accrual software is used to predict the accrual at the end of a clinical trial given accrual to the present time, and can provide the probability that the trial will finish within the planned time frame or the time frame required to recruit the planned number of subjects. The dissertation concludes in Chapter 5 with summary and future studies.

Figure 1-1. A graphical illustration of variance and bias, with (A) estimators are unbiased and have low variance (B) estimators are unbiased with high variance (C) estimators are biased with low variance and (D) estimators are biased with high variance.



CHAPTER TWO:

Expediting Clinical and Translational Research via Bayesian Instrument Development

(Published on *Applied Psychometric Measurement*)

Abstract

Developing valid and reliable instruments is crucial but costly and time-consuming in health care research and evaluation. The Food and Drug Administration and the National Institutes of Health have set up guidelines for developing patient-reported outcome instruments. However, the guidelines are not applicable to cases of small sample sizes. Instead of using an exact estimation procedure to examine psychometric properties, our Bayesian Instrument Development (BID) method integrates expert data and participant data into a single seamless analysis. Using a novel set of priors, we use simulated data to compare BID to classical instrument development procedures and test the stability of BID. To display BID to non-statisticians, a graphical user interface based on R and WINBUGS is developed and demonstrated with data on a small sample of heart failure patients. Costs were saved by eliminating the need for unnecessary continuation of data collection for larger samples as required by the classical instrument development approach.

*Key Words***:** BID, content validity, construct validity, patient reported outcomes, reliability

Introduction

Patient reported outcomes (PRO), which specifically refer to self-reports by patients, play an important role in health care research, system evaluation, and drug approval. Both the National Institutes of Health (NIH; Cella et al., 2010) and the Food and Drug Administration (FDA; Speight & Barendse, 2010) have set up guidelines for developing and accessing patient reported outcome measurement instruments. As implied in the extensive guidelines, classical psychometric instrument development is a lengthy process. Establishing a valid and reliable instrument requires necessary but time-consuming data gathering on content validity and on construct evidence. Typically, researchers begin the instrument development process by first estimating content validity, which includes subject matter experts' evaluations of the extent to which items match the theoretical definition of the construct (Crocker & Algina, 1986). Estimating content validity is based on an interpretive trait-based argument, thereby providing a framework for linking the target domain to items that are developed and revised based on content expert opinions (Polit & Beck, 2006). Expert opinions only are used for making decisions about item contents (i.e., evidence of content validity), such as removing non-relevant items, fine tuning the wording of items, or adding new items. Next, through the collection of data from participants, the reliability and construct validity of items are tested by examining item inter-relationships using factor analysis methods or an item response theory (IRT) model, which is an ordinal version of factor analysis from a modeling standpoint (Beck & Gable, 2001). In classical psychometric instrument development, the information provided by experts and by participants usually is analyzed separately. Information from the expert data (content analysis) is not used in a factor analysis (or IRT) of the participants' data (Messick, 1989).

Classical psychometric instrument development typically requires a large number of subjects (Westland, 2010; Westland, in press). Identifying sufficient numbers of subjects, particularly when there are limited resources and/or small populations, is not optimal for the timely development of needed instruments. For example, the development of an instrument for measuring outcomes for bipolar disorder took four years (Michalak & Murray, 2010). Therefore, a valid method that requires fewer subjects and less time is needed to expedite collecting data for research and translating findings into practice. The FDA and NIH do provide guidelines for factor analysis and IRT; however, they do not address situations for which targeted populations are small, which can arise when evaluating small racial or ethnic groups or patients suffering from rare disease. Lee (1981) developed a Bayesian approach for factor analysis and Lee and Song (2004) showed that it can be used for small sample sizes (N=50) while traditional factor analysis cannot. Gajewski et al. (2013) developed a unified psychometric model that potentially needs fewer subjects in estimating validity evidence for establishing new instruments. The Bayesian approach integrates the analyses of the content and construct validity (IACCV; Gajewski, Price, Coffland, Boyle, & Bott, 2013). In brief, the method first summarizes the state of knowledge of content validity using a *prior* probability distribution derived from expert judgment. The participant data are then used to update to a *posterior* distribution. IACCV was first used on an instrument measuring nursing home culture change (Gajewski, Price, Coffland, Boyle, & Bott, 2013). With this method, the investigators can transfer some of the response burden from the participants to an expert panel resulting in a faster, more efficient, and less costly instrument development process.

In this paper, we further advance IACCV methodology and report an easy-to-use software package that will implement our Bayesian instrument development method targeted for

use by clinical researchers. Among the advancements, some are statistical and some are practical (e.g., software development). For example, in our previous work (Gajewski, Price, Coffland, Boyle, & Bott, 2013), a logit transformation was used for item-domain correlations, which is too restrictive because it produces only non-negative correlations. In the current study, we use Fisher's transformation (Wilks, 1962) because its range (–1 to 1) is a real representation of correlation. To reflect these new advancements in IACCV in the form of our new software for clinical researchers, the current approach is called Bayesian instrument development (BID).

Evaluation and comparison of the Bayesian approach to likelihood methods has been studied by many researchers (Chaloner, 1987; Browne & Drapper, 2006). In general, a Bayes estimator has a smaller squared error but a larger bias compared to a maximum likelihood estimator in analysis of variance components (Chaloner, 1987). Comparing the Bayesian approach to traditional factor analysis was first investigated by Lee (1981) using simulation studies. The results showed that Bayes estimates using "reasonably good prior information" are better than flat priors, and both of the estimates have smaller root mean square error than the traditional maximum likelihood factor analysis approach (Lee & Shi, 2000). Samaniego and Reneau (1994) presented a landmark study which showed both theoretically and empirically that Bayesian estimates perform better than frequentists' when the priors are close enough to the truth. However, when the priors are highly misinformed about parameters, Bayesian estimates may lose their superiority and should be processed with caution . We have extended the study into a comparison of BID and traditional factor analysis for various levels of expertise: some experts have a correct opinion and others have a wrong opinion. The effect of these "contaminated priors" on Bayesian factor analysis and its comparison to a classical approach is a main contribution of this paper.

Using simulated data, we test the BID approach by comparing it with classical instrument development in terms of performance stability and time consumption in development. The results will demonstrate that while the mean square error (MSE) of correlation estimation using classical instrument development does not change with respect to the number of experts, BID has lower error with even a single expert and further improves estimation efficiency as the number of experts increases. We will also demonstrate that the mean squared error for BID is smaller when compared to classical instrument development in the case when experts are biased by as much as 50% and the number of biased experts is small (e.g., 1 or 2 biased experts out of 6 total experts). To make BID user-friendly, we programmed a graphical user interface (GUI) version of BID using R and WinBUGS (R Development Core Team, 2012; Lunn, Thomas, Best, & Spiegelhalter, 2000). The GUI version of BID can guide users to retain widely accepted principles (e.g., reliability, factor structure, or item characteristics) of instrument development by simple point and click. The new BID software is demonstrated to clinical researchers who are non-statisticians and applied to a research project *Timeliness of Symptom Recognition, Interpretation, and Reporting in Heart Failure* (TSRIR) (K99NR012217).

BID model

Bayesian Instrument Development (BID) expands Integrated Analysis of Content and Construct Validity (IACCV) that was first developed by Gajewski et al. (2013). In the current paper, we will focus on a one factor (domain) BID model. For details of the general model, please see Gajewski et al. (2013). Because of its general use in practice and simple analytic form, we motivate the factor analytic model with a simple illustrative calculation of item-domain correlation.

Simple illustration: item to domain correlation

Suppose that the true distribution of item to domain correlation, after Fisher transformation, is normal; that is, $g(r) \sim N(g(\theta), \frac{1}{n})$, where *r* is the sample item to domain Pearson's correlation (e.g., standard calculation in psychometric software), *n* is the number of participants, and *g*(.) is a function, Fisher's transformation. Then, suppose the likelihood of correlation in a transformed scale can be written as $g(r)|g(\rho) \sim N(g(\rho), \frac{1}{n})$. The prior for experts has a normal distribution after Fisher transformation $g(\rho) \sim N(g(\rho_0), \frac{1}{n_0})$, where g($\rho_0$) is the transformed prior mean item to domain correlation and n_0 is the prior sample size. The posterior distribution of g(ρ) is then $g(\rho)|g(r) \sim N(\frac{ng(r)+n_0 g(\rho_0)}{n+n_0}, \frac{1}{n+n_0})$, where the mean is a linear combination of the transformed sample correlation and the prior mean, weighted by their respective sample sizes. The variance is simply the inverse of the sum of the observed and prior sample sizes.. After assuming a true but unknown fixed correlation, θ, we can calculate the mean squared error using the standard formula $MSE(E[g(\rho)]) = Var(E[g(\rho)]) + Bias^2$, where $E[g(\rho)]$ is the Bayes estimate for $g(\rho)$. Thus using standard expected value calculations we can see that $Bias = \frac{ng(\theta)+n_0 g(\rho_0)}{n+n_0} - g(\theta) = \frac{n_0(g(\rho_0)-g(\theta))}{n+n_0}$. Further, using standard variance calculation $Var(E[g(\rho)]) = \frac{n^2}{(n+n_0)^2}\frac{1}{n} = \frac{n}{(n+n_0)^2}$, and so the mean squared error is $MSE(E[g(\rho)]) = Var(E[g(\rho)]) + Bias^2 = \frac{n}{(n+n_0)^2} + \frac{{n_0}^2}{(n+n_0)^2}(g(\rho_0) - g(\theta))^2$. It is easy to show that using the *MSE* that BID is better than classic when $\frac{n}{(n+n_0)^2} + \frac{{n_0}^2}{(n+n_0)^2}(g(\rho_0) - g(\theta))^2 < \frac{1}{n}$, or $(g(\rho_0) - g(\theta))^2 < (\frac{1}{n} - \frac{n}{(n+n_0)^2})\frac{(n+n_0)^2}{{n_0}^2}$. Suppose the prior sample size is n_0=30 and data sample size is *n*=77. Using the inequality, BID is better than classical method when $(g(\rho_0) - g(\theta))^2 < \left(\frac{1}{n} - \frac{n}{(n+n_0)^2}\right)\frac{(n+n_0)^2}{{n_0}^2} = (\frac{1}{77} - \frac{77}{(77+30)^2})\frac{(77+30)^2}{30^2}$=0.079654.

Missing from this illustrative example is a methodology for specifying the prior distribution as well as specification to a more general model like factor analysis. We explore these issues in the following sections and develop an estimation approach faster than Gajewski et al. (2013) for our simulations studies to calculate the mean squared error for a comparison of BID and the classical approach.

Model and computation for experts

For classical content validity, content experts are instructed to review all items and to "…indicate how relevant you perceive the item to be to the respective domain." For each item, the rating scale generally ranges from 1 indicating 'not relevant' to 4 signifying 'highly relevant'. A recent study showed that content validity can also be measured using correlation scales, which are equivalent to relevance scales (Gajewski et al., 2012). Therefore, in addition to relevancy, content validity can be further interpreted as the experts' opinion regarding item-to-domain correlation.

For the content expert data, suppose there are K experts ($k = 1, 2, 3, \ldots, K$) responding to p questions ($j = 1, 2, 3, \ldots, p$) and let x_{jk} represent the ordinal measure of the k^{th} expert's opinion for the j^{th} item's relevance. The correlation for the j^{th} item with its respective domain as perceived by the k^{th} expert is denoted by ρ_{jk} and its assumed mapping from x_{jk} is as follows:

$$x_{jk} = \begin{Bmatrix} \text{'not relevant'} & \text{if} & 0.00 \le \rho_{jk} < 0.10 \\ \text{'somewhat relevant'} & \text{if} & 0.10 \le \rho_{jk} < 0.30 \\ \text{'quite relevant'} & \text{if} & 0.30 \le \rho_{jk} < 0.50 \\ \text{'highly relevant'} & \text{if} & 0.50 \le \rho_{jk} \le 1.00 \end{Bmatrix} \quad (1)$$

The correlation between the item and domain after pooling information from all experts is expressed as $\rho_j = corr(f, z_j)$, where f, the factor score of the domain, has a standard normal distribution with mean of zero and variance of one, and z_j is the standardized response for item j.

The thresholds of ρ_j are argued under the assumption that the experts view correlations using Cohen's cut points (Cohen, 1988). Equation (1) allows us to properly model the uncertainty in these correlations that is induced by the interval censoring (e.g., < 0.10, < 0.30, etc.), similar to other methods used in studies on pressure ulcer staging (Gajewski, Hart, Bergquist, & Dunton, 2007), hearing test for audiology (Gajewski, Sedwick, & Antonelli, 2004; Gajewski, Nicholson, & Widen, 2009), and instrument development measuring nursing home quality of life (Gajewski, Price, Coffland, Boyle, & Bott, 2013). Under some circumstances, if the prior data are from a panel knowledgeable in the area but not considered experts, an equal space transformation is found to be more appropriate and closer to actual correlations than the unequally spaced transformation (Gajewski et al., 2012; Pawlowicz et al., 2012). Equation (2) shows the equally spaced model:

$$x_{jk} = \begin{cases} \text{'not relevant'} & \text{if} \quad 0.00 \le \rho_{jk} \le 0.25 \\ \text{'somewhat relevant'} & \text{if} \quad 0.25 < \rho_{jk} \le 0.50 \\ \text{'quite relevant'} & \text{if} \quad 0.50 < \rho_{jk} \le 0.75 \\ \text{'highly relevant'} & \text{if} \quad 0.75 < \rho_{jk} \le 1.00 \end{cases} \tag{2}$$

Using Fisher's transformation, ρ_j is transformed to

$$g(\rho_j) = \frac{1}{2} \log \frac{1+\rho_j}{1-\rho_j} \tag{3}$$

A hierarchical model that combines experts and includes all items is:

$$g(\rho_{jk}) = g(\rho_j) + e_{jk}, \tag{4}$$

where e_{jk} is the random error assuming e_{jk}~i.i.d. $N(0,\sigma^2)$.

Assuming a lack of information on content validity prior to eliciting expert judgment, we use an essentially flat prior σ^2~IG (.00001, .00001) and $g(\rho_j)$~N(0, 3^2). The shape and rate parameters for inverse gamma (IG) are both 0.00001. The prior specification allows us to

estimate the posterior distribution of ρ_j using the expert's data via a Markov chain Monte Carlo (MCMC) procedure implemented in the software WinBUGS. These results then form the prior distribution of ρ_j and are combined with the likelihood generated by participant data in the construct validity and reliability analysis. Note that this is a pool of expert's data for a simple, compromised prior, which is different from Samaniego & Reneau (1994), who analyze single expert at a time.

The previous model (Gajewski, Price, Coffland, Boyle, & Bott, 2013) for expert data utilized a logit transformation of correlation. However, because its range is 0 to 1 it can only be applied after reverse scoring items that are negatively correlated with the domain. For this reason, the logit transformation might be too restrictive. Using Fisher's transformation in the current model allows a real line representation of correlation from −1 to 1.

Based on our previous studies for nursing home culture change and nurse job satisfaction, we have found that the distribution of $g(\rho_j)$ can be approximated normal. The variance of $g(\rho_j)$ has an inverse relationship with the number of experts. That is, it is a simple representation of the sample size for the Bayesian prior. Specifically, the distribution of $g(\rho_j)$ can be written as

$$g\left(\rho_j\right) \sim N\left(g\left(\rho_{0j}\right), \frac{1}{n_{0j}}\right) \tag{5}$$

where $n_{0j} = 5 \times K$ and K is the total number of experts. Thus, the above equation is used in the simulation study. And it assumes each expert is worth five participants. The calculation of the variance and the justification of this estimation can be found in the Appendix. This approximation is not used in the actual BID software, but rather is used to study the one-factor analytic model's properties via simulation.

Model and computation for participants

For construct validity, we use a standardized factor analytic model to describe the item-to-domain correlations for participants' data (Gajewski, Price, Coffland, Boyle, & Bott, 2013). Specifically, the one factor (domain) model is:

$$z_{ij} = \rho_j f_i + e_{ij}. \quad (6)$$

Assuming y_{ij} is the i^{th} subject's response to the j^{th} item, it is standardized to z_{ij}, which allows us to focus on estimating item-to-domain correlations. The total number of participants is *n* and the total number of items is *p*; f_i is the factor score for the domain (e.g., overall pain); and is typically assumed to follow a standard normal distribution, *N* (0, 1). The correlation between z_{ij} and f_i is represented by ρ_j; thus, ρ_j^2 is the "reliability" of item *j* ((Alonso, Laenen, Molenberghs, Geys, & Vangeneugden, 2010). The assumption $z_{ij} \sim N(0, 1)$ implies $e_{ij} \sim N(0, 1-\rho_j^2)$. Thus, the likelihood for z_{ij} is:

$$L(\mathbf{z} \mid \boldsymbol{\rho}, \mathbf{f}) = \prod_{i=1}^{n} \prod_{j=1}^{p} N\left(z_{ij} \mid \rho_j f_i, 1-\rho_j^2\right). \quad (7)$$

As discussed previously, the distribution of $g(\rho_j)$ is approximately normal $g(\rho_j) \sim N(g(\rho_{j0}), \sigma^2_{g(\rho_{j0})})$. To estimate the posterior distribution of the j^{th} component of **ρ,** a flat prior N(0, 3^2), representing high prior uncertainty in the construct validity of items can be used. An informative prior can also be utilized by obtaining the mean and variance of $g(\rho_{j0})$ from content analysis using method of moments. In our simulation study, $g(\rho_j) \sim N\left(g(\rho_{0j}), \frac{1}{n_{0j}}\right)$ is used for the informative prior. Using this prior, the posterior distribution is:

$$\pi(\boldsymbol{\rho}, f \mid n_0, z) \propto$$
$$\prod_{i=1}^{n}\prod_{j=1}^{p}\{N(Z_{ij} \mid \rho_j f_i, 1-\rho_j^2)\} \times \prod_{i=n}^{n} N(f_i \mid 0,1) \times \prod_{j=1}^{p} N(g(\rho_j) \mid g(\rho_{j0}), \frac{1}{n_{j0}}) / \left((1+\rho_j)(1-\rho_j)\right). \quad (8)$$

The estimation of the posterior distribution of **ρ** can be obtained by MCMC methods performed in R and WinBUGS. However, this process is unnecessarily slow for simulation tests where only the posterior mode is needed. For example, when the number of items is 8 and the sample size is 100, it takes about 10 hours to compute 1,000 iterations using Equation (8) on a shared high speed workstation with dual Xeon 3.6 GHz processors and 4 GB of SDRAM.

Equation (8) provides the full posterior distribution that we use for clinical research, as it contains all the information for our inference. Although the posterior distribution does not have a closed form solution, it can be simplified because $\boldsymbol{z_i}$ follows a multivariate normal distribution, $\boldsymbol{z_i} \sim MVN(0, \boldsymbol{R})$. The variance of $\boldsymbol{z}_i$ is 1, the covariance between z_{ij} and z_{ik} is the product of ρ_i and ρ_j, and $\boldsymbol{R}$ is

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_1\rho_2 & \cdots & \rho_1\rho_p \\ \rho_2\rho_1 & 1 & \cdots & \rho_2\rho_p \\ \vdots & \vdots & \ddots & \vdots \\ \rho_p\rho_1 & \rho_p\rho_2 & \cdots & 1 \end{bmatrix}. \quad (9)$$

Then, the posterior distribution can be simplified as:

$$\pi(\boldsymbol{\rho}, \boldsymbol{f} \mid n_0, \boldsymbol{z}) \propto \prod_{i=1}^{n} MVN(\boldsymbol{z}_i \mid 0, \boldsymbol{R}) \times \prod_{j=1}^{p} N(g(\rho_j) \mid g(\rho_{j0}), \frac{1}{n_{j0}}) / \left((1+\rho_j)(1-\rho_j)\right). \quad (10)$$

The specific estimation of ρ_j's can be obtained by calculating the posterior mode of (10), which is the minimum of the following formula:

$$-\log(\pi(\boldsymbol{\rho} \mid n_0, \boldsymbol{z})) =$$
$$-\sum_{i=1}^{n} MVN(\boldsymbol{z}_i \mid 0, \boldsymbol{R}) - \sum_{j=1}^{p} \log(N(g(\rho_j) \mid g(\rho_{j0}), \frac{1}{n_{j0}})) + \log(1+\rho_j) + \log(1-\rho_j) + C' \quad (11)$$

with constant, C. The estimation of ρ_j's can be obtained by using the "nlminb" function in R, which is an optimization procedure using PORT routines (Bates & Sarkar, 2012). It has been shown in the literature (Lee, 2007) that using multivariate-based equations, like Equation (11), for estimating the *full posterior distribution* is *slower* than using univariate-based equations, like Equation (8). However, we only need the *posterior mode* for our simulation study, so using Equation (11) is actually much faster and more convenient. For example, the processing time is *reduced* to about 5 minutes for the same analysis as described above using Equation (8).

Calculation of item reliability

In addition to validity, reliability is also an important concept in instrument development because it represents the precision of measurements. Cronbach's alpha has been widely used in the estimation of reliability. However, it is a lower bound estimator (Green & Yang, 2009), and therefore a conservative assessment of reliability. It is better to estimate composite reliability via a model based approach (Alonso, Laenen, Molenberghs, Geys, & Vangeneugden, 2010).

Under the one factor model, equation (6) also can be written in matrix form: $\boldsymbol{z}_i = \boldsymbol{\rho} \boldsymbol{f}_i + \boldsymbol{e}_i$

where, $\boldsymbol{z}_i' = \left(z_{i1}, z_{i2}, \cdots, z_{ip} \right)$ and $\boldsymbol{\rho}' = \left(\rho_1, \rho_2, \cdots, \rho_p \right)$.

Because $e_{ij} \sim N(0,\ 1-\rho_j^{\ 2})$, the residual covariance matrix for $\boldsymbol{e}_i$ can be written as:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1-\rho_1^1 & 0 & \ldots & 0 \\ 0 & 1-\rho_2^2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1-\rho_p^2 \end{bmatrix}. \qquad (12)$$

According to Alonso et al. (2010), the entire reliability is:

$$\boldsymbol{R}_{\Lambda} = 1 - \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{R}|}, \qquad (13)$$

where $\boldsymbol{z}_i \sim MVN(0, \boldsymbol{R})$, and $\boldsymbol{e}_i \sim MVN(0, \boldsymbol{\Sigma})$. Entire reliability measures the reliability of the whole domain of items. Our extension using BID allows for efficient estimation of reliability for small populations.

Simulation study

We use simulated data to evaluate the performance of BID and compare it to the traditional factor analysis, a common approach for validating statistical methods. In BID, we assume all experts agree in interpreting the notion of correlation in their judgment response. Furthermore, experts' data are correlated with participants' data, indicating their judgments are the same or highly similar. To understand how violations of these assumptions may affect the analysis of data from expert opinion, we will test violations of the assumptions about performance of the experts. Using simulation, we will see more clearly the benefit-to-cost ratio of the number of experts needed for adequate evidence of content validity, and it will allow for exploration of the sensitivity of our assumptions regarding the experts' judgments.

According to our previous job satisfaction study, the average of the correlations ρ between individual item and domain is 0.5 (Gajewski, et. al., 2012); therefore, we will assume the true correlation is $\rho_{TRUE} = 0.50$ for all items. Under this assumption the scale score would explain 25% of the variation in the corresponding item. We obtain the Bayesian estimate—the posterior mode—using R instead of using the MCMC procedure in R and WINBUGS that estimates the full posterior distribution. The latter is used in our R-GUI BID software developed for clinical study researchers (see section 4). For comparison to traditional methods, classical factor analysis is done using the "Lavaan" package built in R (Rosseel, 2012). The numerical results of the "Lavaan" package are typically very close to that of the commercial package Mplus (Rosseel, 2010).

We test BID under two scenarios. First, we examine the stability of BID using different assumptions (unbiased experts ρ_0=0.50 or biased experts ρ_0=0.75), as the experts tend to overestimate the relevance of items in content validity (Gajewski, Price, Coffland, Boyle, & Bott, 2013). Specifically, we assume the prior distributions are $g\left(\rho_j\right) \sim N\left(g\left(0.5\right), n_{0j}^{-1}\right)$ or $g\left(\rho_j\right) \sim N\left(g\left(0.75\right), n_{0j}^{-1}\right)$. We then test BID performance when prior distributions (experts) are a mixture of unbiased and biased information.

Biased or unbiased priors

In order to test the performance of BID under biased prior expert information, we use a four factor simulation design. The first simulation factor is the presence or absence of expert bias—either (1) the judgment of correlation between experts provides estimates of the square root of item reliability (ρ_0=0.50, unbiased for all items), or (2) the experts are biased (ρ_0=0.75, for all items). We also examine the impact of bias over a range of number of items (8, 16, or 24), number of subjects (50, 100, or 200) and number of experts (1, 2, 6, or 16), which are typical instrument development studies with small sample sizes (Polit & Beck, 2006).

We use R to generate participant data based on the assumption that the standardized item response z_i is multivariate normal with mean vector zero and variance of identity. Each factor combination simulation is repeated for 1,000 times. Let $\hat{\rho}_j(s)$ be the posterior mode for the s^{th} simulation in a total of 1,000 iterations. Then $MSE_j = \sum_{s=1}^{1000}(\hat{\rho}_j(s) - \rho_j)^2 \Big/ 1000$ and $\overline{MSE} = p^{-1}\sum_j MSE_j$; $\boldsymbol{Bias^2}_j = \sum_{j=1}^{1000}\left(\hat{\boldsymbol{\rho}}_{\boldsymbol{j}}(\boldsymbol{s}) - \boldsymbol{\rho}_{\boldsymbol{TRUE}}\right)^2 / 1000$; and $\overline{\text{Bias}^2} = p^{-1}\sum_j Bias^2{}_j$ .In our study, the $\overline{MSE}$ and $\overline{\text{Bias}^2}$ are used for comparison via graphs.

Figure 2-1 shows the MSE correlation estimates for unbiased and biased experts with participant sample sizes of 50, 100, or 200 when the number of items is 8. As traditional factor analysis does not use content validity information, it is not surprising that the MSE is unaffected by prior information. When the priors are unbiased, the MSE obtained by BID is always smaller than classical factor analysis when the sample size is 50 or 100, indicating BID is more efficient. When the sample size is large (n=200), the MSE by classical factor analysis is slightly smaller than that of BID only when the number of experts is one. As the number of experts increases, the gains in efficiency by BID also increase. However, when the priors are biased, the relative efficiency of BID compared to classical factor analysis is a function of the number of experts. When the number of experts is small (e.g., one or two), BID has lower MSE than traditional methods. As the number of experts increases, any gains in efficiency recede and eventually traditional methods become superior. When sample size is smaller (e.g., n=50), efficiencies gained by BID are most obvious. The differences in MSE when using BID versus classical factor analysis follow a similar pattern when the number of items is 16 or 32. The corresponding figures are shown in the supplemental sections.

Overall, our first simulation study shows the impact of biased experts on the efficiency of BID relative to traditional methods is a function of the number of biased experts. When the expert data are unbiased, the MSE is always smaller when using BID rather than classical factor analysis for samll sample studies. However, a stubbornly biased subjective prior is quite harmful. Our findings are consistent with those discussed in other Bayesian literature investigating Bayesian and classical estimation (Samaniego and Reneau 1994, Garthwaite, Kadane, & O'Hagan, 2005).

Contaminated priors

A second simulation study assessed the impact of having a contaminated expert panel; that is, assuming the number of experts is six and we investigated the impact of having one or two with biased judgment. Here, we let *c* represent the number of biased experts and examine the effects of prior contamination for participant sample sizes of 50, 100, and 200; and numbers of items of 8, 16, and 24.

We investigated the theoretical distribution of $g(\rho)$ when the total number of experts is 6, and the number of biased experts (ρ_0=0.75) is c = 0, 1, 2, or 6, respectively. When all experts are unbiased (ρ_0=0.50), the distribution of $g(\rho)$ is normal with mean $0.5\log(3)$ and variance $1/30$. The distribution of $g(\rho)$ shifts to the right when there is one biased expert in the panel. The upper tail becomes heavier when contaminated priors increase to two. When all the experts are biased, the distribution of $g(\rho)$ changes back to a normal distribution with mean $0.5\log(7)$, and variance 1/30. The theoretical distributions of $g(\rho)$ and ρ are shown in supplementary figures 2-S1A and 2-S1B.

Similar to the first set of simulations, R was used to generate data and calculate the posterior mode for $\boldsymbol{\rho}$, and 1,000 iterations were carried out for each simulation scenario. The MSE shown in the graph is $\overline{MSE}$ from the 1,000 replications.

As shown in Figure 2-2A, compared to classical factor analysis, BID always has a lower MSE of correlation estimates when the number of biased experts is one or two for a panel with six experts. Fewer contaminated priors are always better, as the data shows the MSE with only one biased expert is lower than that of two biased experts. The results are very similar across participant sample sizes of 50, 100 and 200, but the gain in efficiency decreases as the number of participants increase. Our results show that when the number of items is 8, the MSE of BID with one contaminated priors is equivalent to the MSE from a classical CFA method (MSE = 0.01).

However, the level of precision is achievable with half the number of participants (n = 100 for CFA, n = 50 for BID). Figure 2-2B shows the bias^2 of correlation estimates. The square of bias is larger using BID when there are biased experts in the panel compared to the classical method. The results of bias^2 indicated that BID lowers MSE through decrease of variance. Overall, estimation using informative priors is more efficient, which confirms the novel finding of Samaniego and Reneau (1994).

Application to data from heart failure study and development of GUI-BID software

We programmed a graphical user interface (GUI) version of BID using "TclTk" and "fgui" package in R and WinBUGS (Grosjean, 2012; Thomas & Nan, 2009). The new user-friendly BID software was readily usable by clinical researchers who were non-statisticians. An example of the GUI-BID software window can be found in figure 2S-4. The clinical researchers can choose to use a flat prior or an informative prior from expert data by selecting the option "prior" to be true or false. Data loading for both expert and participant data is accomplished by using the BID software in a point and click environment. The users are also guided to add the factor structure for the instrument. Although we focused on a one factor model for this paper, the software applies to multi-factor models as well.

In a study of heart failure, we apply BID to develop an instrument with the purpose to assess symptoms in heart failure patients. Using the proposed instrument, *Timeliness of Symptom Recognition, Interpretation, and Reporting in Heart Failure (TSRIR)* (K99NR012217), the study is conducted by nursing researchers to examine the symptom experiences of patients with heart failure leading up to hospitalization for acute decompensated heart failure. The tested instrument has ten items, and the experts' ratings are on a 1-4 scale, from "not relevant" to "highly relevant".

First, the experts' data are analyzed using the Bayesian approach with equal space transformation discussed in section 3. In the MCMC procedure for estimating ρ_j we used three chains and a burn-in of the first 2,000 draws of each chain. The next 10,000 iterations were used for inferences. We also perform BID on participant data using a flat prior; that is, using $g(\rho_j) \sim N(0,3^2)$ in the BID model. The results are shown in Table 2-S1. The entire reliability using the flat prior specification is 0.943, and its standard deviation is 0.010. Using an informative prior, $g(\rho_j) \sim N(g(\rho_{j0}), \sigma^2_{g(\rho_{j0})})$, with an equal spaced model, the total reliability is 0.938 with standard deviation equal to 0.007. According to NIH PROMIS cut points, the reliability is larger than 0.8 and should be considered substantial (Cella et al., 2010). Overall, in the heart failure study the dimensions (scales) were deemed reliable with a small sample size (n = 60) using BID.

Discussion

Classical instrument development is a powerful tool in patient reported outcome measurements when one has easy access to a large number of participants and considerable resources. However, in practice it usually happens that either the numbers of participants or resources are limited. Using a Bayesian approach, BID provides an innovative method for instrument development. BID is particularly valuable for studies with limited participants, such as studies involving minority groups or those with orphan or rare diseases. In the heart failure study, it is notable that it took over two years to obtain data from 60 patients. The ratio of the variance of reliability by flat prior to informative prior is $0.01^2/0.007^2=2.04$. Thus, 104% more data or 123 participants would be needed if using traditional approaches. Consequently, it took two years to gain this clinical data, and it would take at least two more years without the informative prior. Using the BID approach, the clinical researchers do not need to continue data collection to obtain a larger sample that would be required by the classical instrument

development approach. Both cost and time savings are achieved. Even when large samples are available, limited resources still may present challenges for investigators. For example, the development of an instrument measuring smoking cessation practices in methadone treatment clinics took four years because funds were limited for both participant incentives and personnel. If using BID, the processes could speed up dramatically (Cupertino et al., in review). Overall, BID proposes a psychometrically comprehensive and statistically efficient method for instrument development.

Using indirect measurement information is an important trend in statistics and related areas (Efron 2010). Samaniego and Reneau (1994) revealed that a good utilization of this indirect information, in the form of a prior distribution, can lead to a Bayesian estimator which is superior to a frequentist estimator in most situations. We extended their study into factor analysis for estimation of construct validity. In the current paper, we provided a full comparison of Bayesian approach and traditional factor analysis using simulation studies. In this paper, we investigated the performance of the Bayesian approach under different conditions: (1) the priors (experts) are all unbiased, (2) the priors (experts) are all biased (3) the priors are contaminated, with 1 or 2 biased experts among a total of 6 experts. Our results are consistent with what was in Lee (1981, Lee & Shi 2000), which shows that the Bayesian approach performs better than traditional factor analysis when using a flat or "good informative" prior. Our results are consistent with Samaniego and Reneau (1994)'s general finding that a Bayesian estimate is always superior to a frequentist estimate if the bias in the Bayesian prior is smaller than the sample standard deviation. In addition, we showed more clearly that the Bayesian approach is practically better than traditional methods, especially for small sample sizes as in our heart failure example.

Although the current study focuses on a one factor model, BID can easily be applied to multi-factor models. For the heart failure study data, we also tried different factor structures (i.e., using multiple factors and/or fewer items). The results were consistent with the expected results. In the BID model, we assume the experts' opinion on items is independent and our prior is weighted over all the experts. The assumptions were used to simplify the model. However, in real circumstances the experts' opinions on items are most likely correlated. In future studies, a correlated model may provide a better estimation for the content validity. A hierarchical model could be used to allow for individualized priors for each expert based on their prior beliefs. This would provide us with the ability to evaluate the effect of a single expert on the analysis and draw conclusions, using an approach similar as discussed by Samaniego & Reneau (1994). In BID, we also assume the participants' data are continuous; however, many clinical questionnaires are in ordinal or binary form. Therefore, Ordinal Bayesian Instrument Development (OBID) using item response theory is a topic for future work.

Acknowledgements

Thanks to Drs. K Reeder and Carol Smith for use of their data from grant K99NR012217. This work was supported by a CTSA grant from NCRR and NCATS awarded to the University of Kansas Medical Center for Frontiers: The Heartland Institute for Clinical and Translational Research # UL1TR000001 (formerly #UL1RR033179). The data described is part of a clinical trial supported by the National Institute of Heart, Lung and Blood (NHLBI) Grant # HL 085397, C. Smith RN PhD, PI. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH, NCRR, or NCATS. The first author was partly supported by P20MD004805 Center for the American Indian Community Health (CAICH).

Figure 2-1. Mean Square Error correlation estimates ρ in stimulation study using BID (dashed line) and classical factor analysis (solid line) when the number of items is 8. The number of experts are 1, 2, 6, and 16, and they are either unbiased (ρ_0=0.5, left panel) or biased (ρ_0=0.75, right panel). The participant sample sizes are 50 (A and B), 100 (C and D), and 200 (E and F).

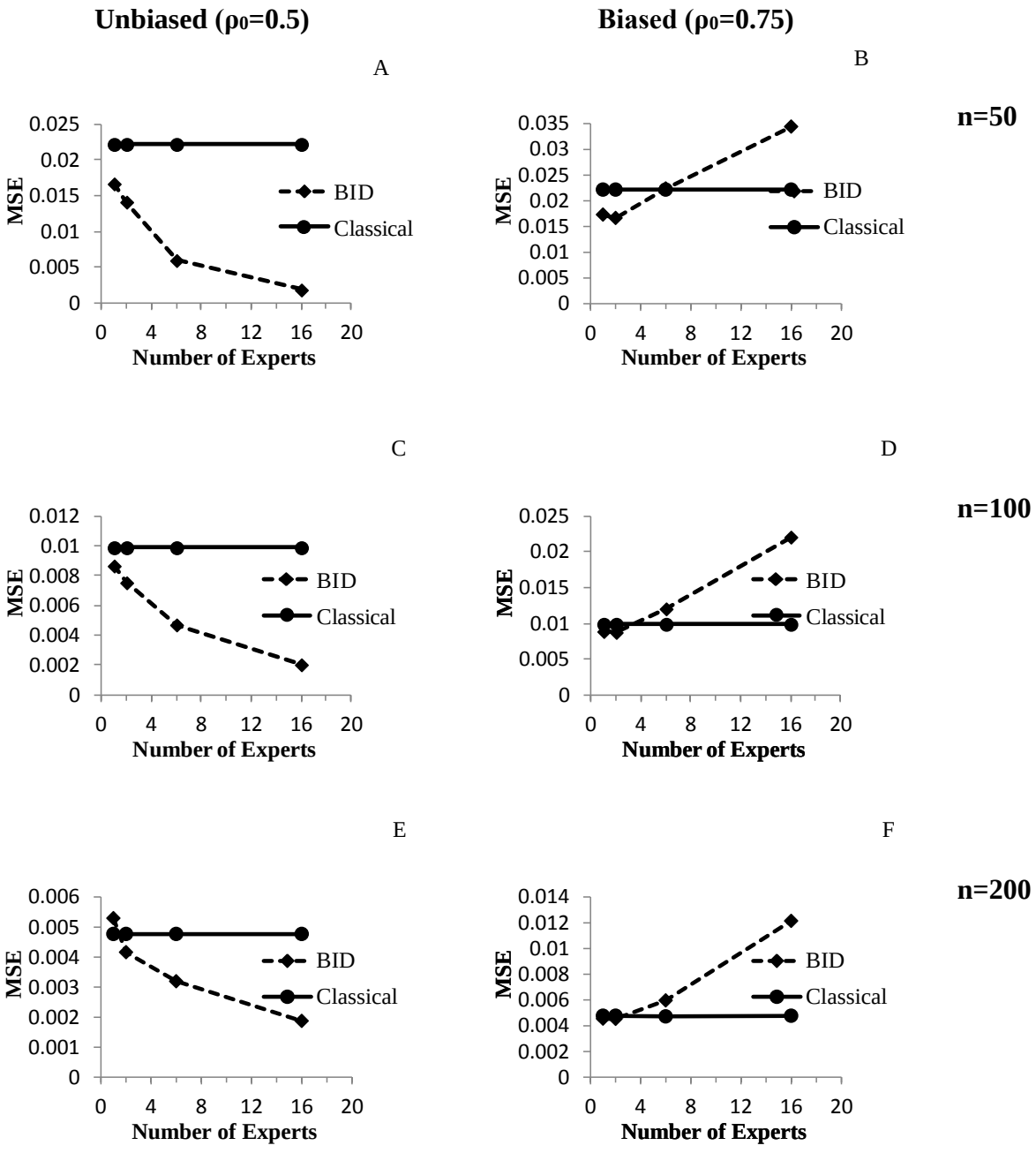
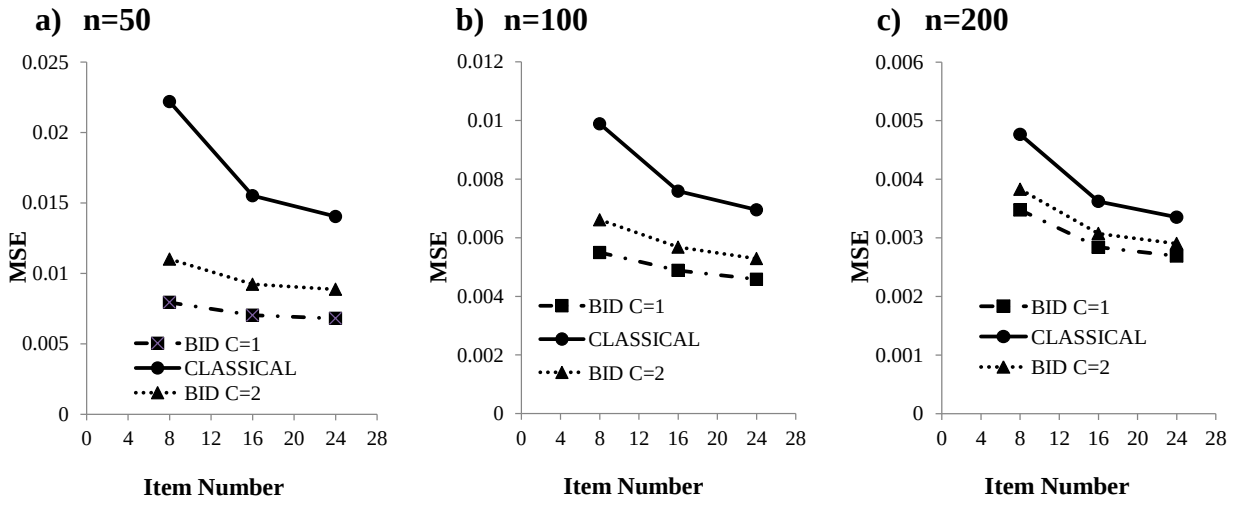


Figure 2-2. Mean Square Error (A) and Square of bias (B) of correlation estimates ρ in stimulation study using classical factor analysis, BID modeling with a panel of six experts containing one biased prior (C = 1), and BID modeling with a panel of six experts containing two biased priors (C = 2) when participant sample size is 50 (a), 100 (b) and 200 (c). All scenarios assume a true correlation ρ_0=0.5 with biased expert judgment of ρ_0=0.75.

A



B

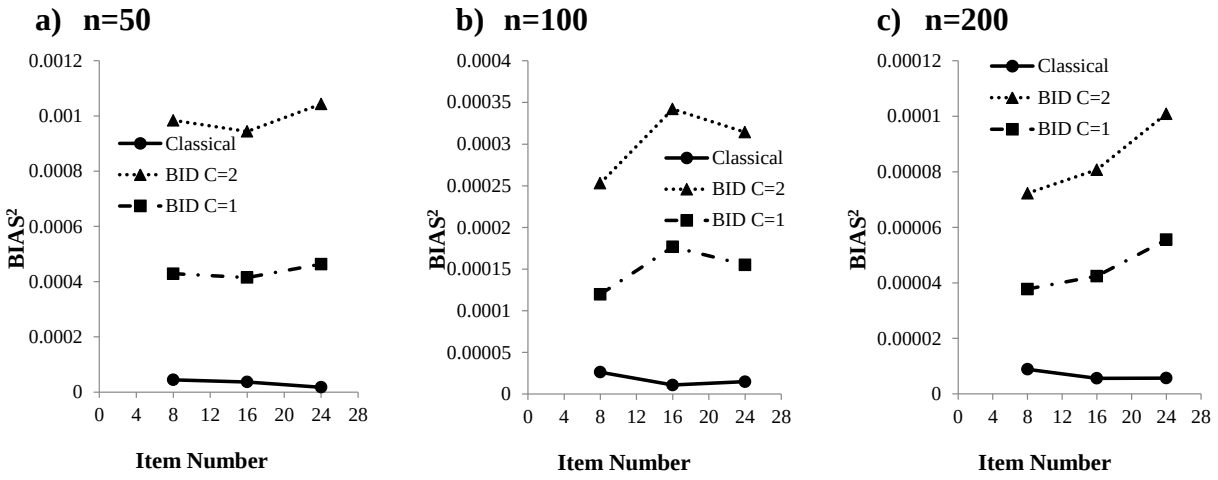


Table 2-S1. The posterior means, standard deviations and 2.5, 50, 97.5 percentage quintiles for the participant data analyzed using BID with flat priors (A) and BID with informative priors using equal space transformation (B).

A) Flat prior

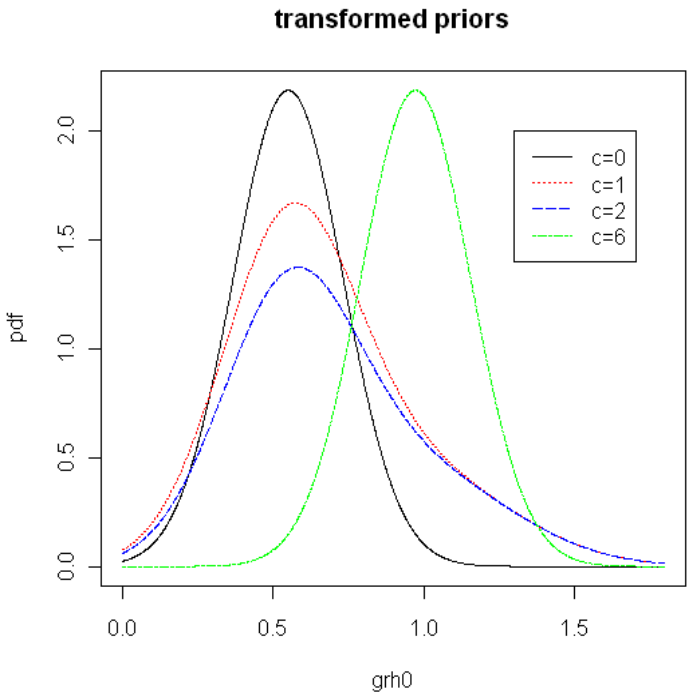
	mean	sd	2.5%	50%	97.5%
item1	-0.034	0.132	-0.286	-0.016	0.195
item2	0.176	0.124	-0.049	0.172	0.413
Item3	0.364	0.126	0.090	0.377	0.587
Item4	0.432	0.115	0.152	0.455	0.596
Item5	0.177	0.162	-0.142	0.173	0.487
Item6	0.920	0.028	0.848	0.925	0.964
Item7	0.829	0.041	0.743	0.834	0.901
Item8	0.927	0.025	0.862	0.931	0.961
Item9	0.455	0.104	0.226	0.461	0.633
Item10	0.532	0.093	0.322	0.537	0.690
$R_{\wedge}$	0.943	0.010	0.921	0.944	0.961

B) Informative prior ("Equal" spaced)

	mean	sd	2.5%	50%	97.5%
item1	0.493	0.063	0.357	0.499	0.600
item2	0.578	0.053	0.471	0.582	0.675
Item3	0.713	0.040	0.628	0.714	0.786
Item4	0.668	0.050	0.559	0.675	0.746
Item5	0.601	0.059	0.471	0.602	0.711
Item6	0.879	0.025	0.819	0.881	0.922
Item7	0.857	0.025	0.803	0.859	0.901
Item8	0.881	0.021	0.832	0.884	0.914
Item9	0.637	0.052	0.532	0.638	0.729
Item10	0.663	0.050	0.549	0.665	0.749
$R_{\wedge}$	0.938	0.007	0.925	0.937	0.948

Figure 2-S1 Distribution of (A) transformed priors g(ρ) and (B) untransformed priors ρ when the total number of experts is 6 for all unbiased experts (c = 0), one biased expert (c = 1), two biased experts (c = 2), and all biased experts ρ_0=0.75 (c = 6). All scenarios assume a true correlation (unbiased) ρ_0=0.5 with biased expert judgment of ρ_0=0.75.

(A)



B)

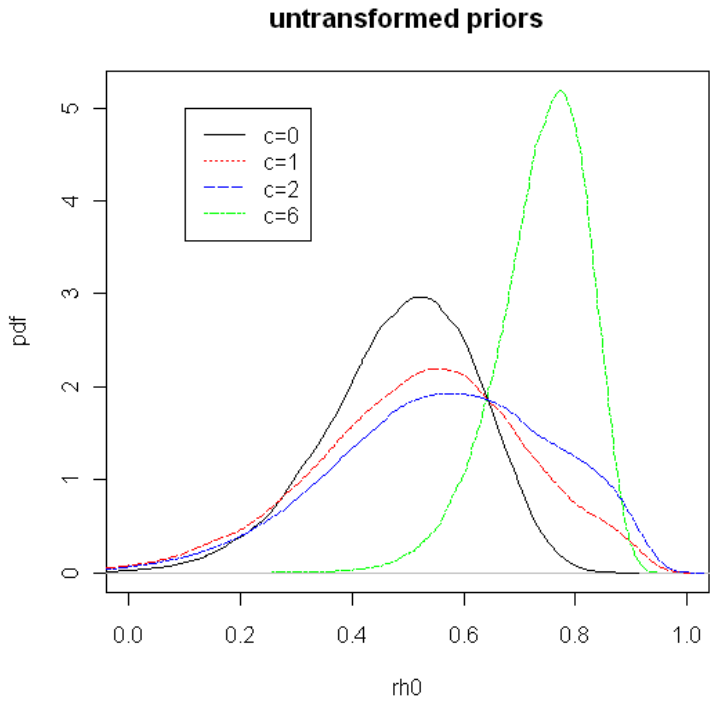


Figure 2-S2. Mean Square Error correlation estimates ρ in stimulation study using BID (dashed line) and classical factor analysis (solid line) when the number of items is 16. The number of experts are 1, 2, 6, and 16, and they are either unbiased (ρ_0=0.5, left panel) or biased (ρ_0=0.75, right panel). The participant sample sizes are 50 (A and B), 100 (C and D), and 200 (E and F).

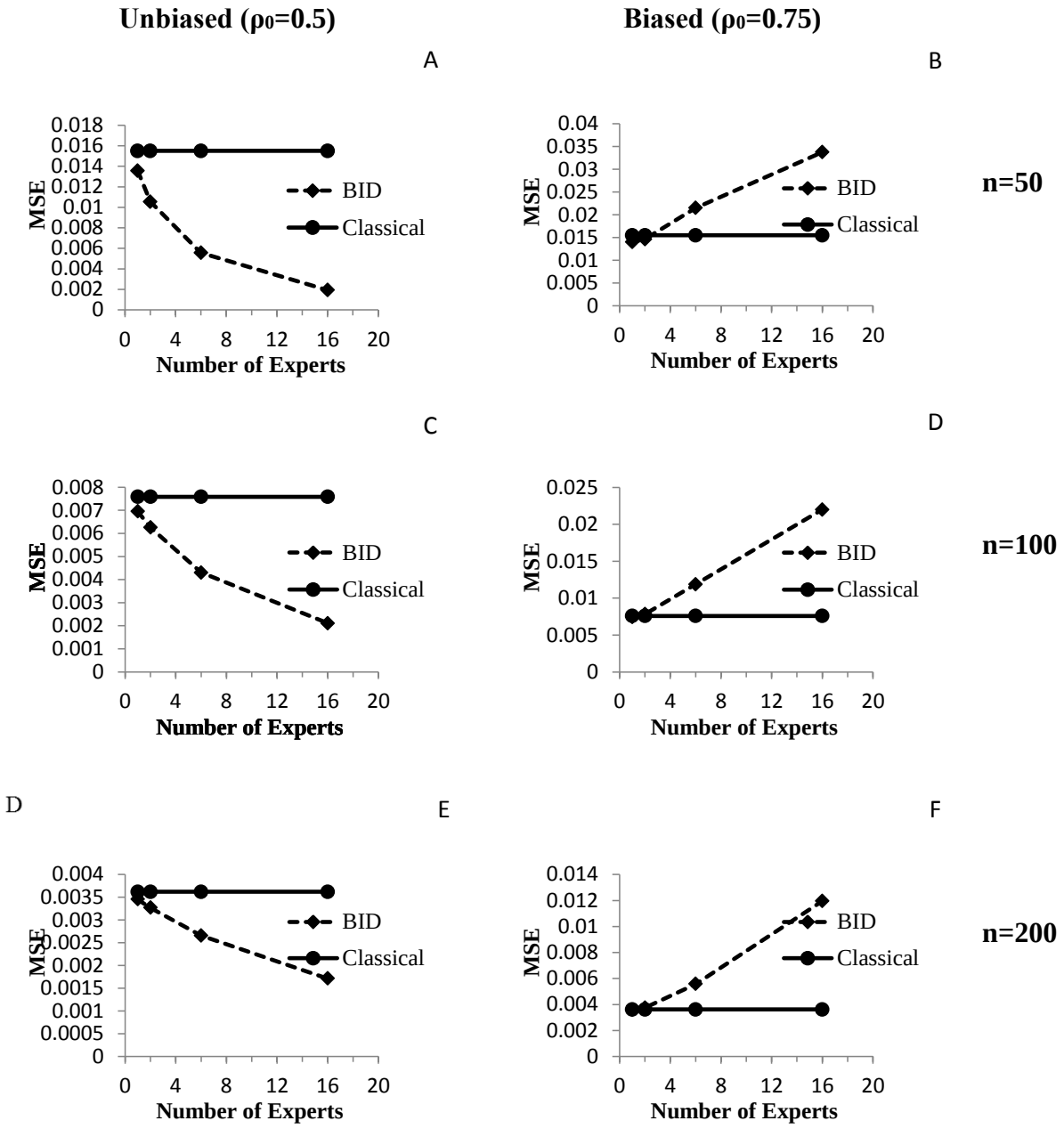


Figure 2-S3. Mean Square Error correlation estimates ρ in stimulation study using BID (dashed line) and classical factor analysis (solid line) when the number of items is 24. The number of experts are 1, 2, 6, and 16, and they are either unbiased (ρ_0=0.5, left panel) or biased (ρ_0=0.75, right panel). The participant sample sizes are 50 (A and B), 100 (C and D), and 200 (E and F).

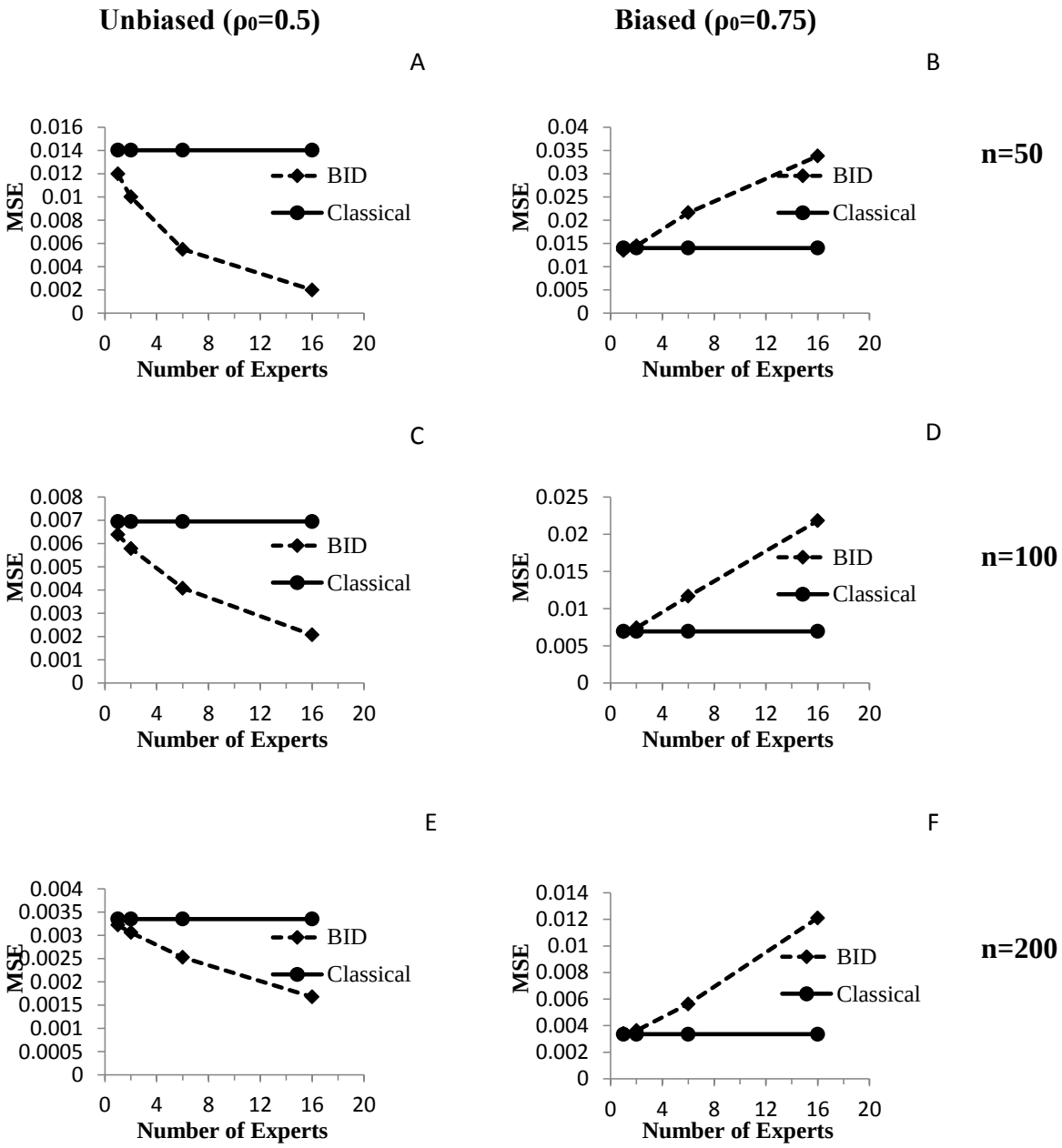
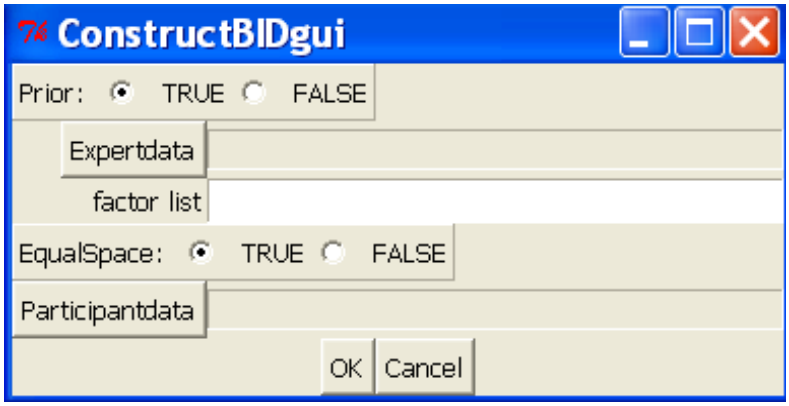


Figure 2-S4. A typical GUI-BID window that guides the clinical researchers analyzing data using BID.



Appendix. Justification of experts equivalents

BID has been used in the study of nursing home culture change (Gajewski et al., 2011) and job satisfaction in nursing faculty members (Pawlowiz et al., 2012; Gajewski et al., 2012). Both studies report the estimation of the correlation of item-to-domain; that is, ρ_j. Using Fisher's transformation:

$$g(\rho_j) = \frac{1}{2}\log\frac{1+\rho_j}{1-\rho_j} \tag{A1}$$

If the variance of ρ_j is V, then by the delta method, the variance of $g(\rho_j)$ is $\left(g'\left(E\left(\rho_j\right)\right)\right)^2 V$. Specifically,

$$g'(\rho_j) = \left(\frac{1}{2}\log\frac{1+\rho_j}{1-\rho_j}\right)' = \frac{\rho_j}{1-\rho_j^2} \tag{A2}$$

Then, the variance of $g(\rho_j)$ can be calculated and the expert equivalent is estimated to be five.

CHAPTER THREE:

Modeling and Validating Bayesian Accrual Models on Clinical Data and Simulations Using Adaptive priors

(Resubmitted to *Statistics in Medicine*.

***The references are formatted for APA style*)**

Modeling and Validating Bayesian Accrual Models on Clinical Data and Simulations Using Adaptive priors

Yu Jiang[1], Steve Simon[4,5] , Matthew S. Mayo[1,3] Byron J. Gajewski[1,2,3] *

1. Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS, 66160

2. School of Nursing, University of Kansas Medical Center, Kansas City, KS 66160

3. The University of Kansas Cancer Center, Kansas City, KS, 66160

4. P.Mean Consulting, Leawood, KS, 66224

5. Department of Biomedical and Health Informatics, University of Missouri - Kansas City, Kansas City, MO, 64110

*Correspondence: Department of Biostatistics, University of Kansas Medical Center, Mail Stop 1026, 3901 Rainbow Blvd., Kansas City, KS 66160 USA. Phone: 913-588-1603, Fax 913-588-0252, Email: bgajewski@kumc.edu

Short Title: Constant Accrual Model on Clinical Trials Using Adaptive Priors

Submitted to *Statistics in Medicine*, 4/30/2014

Resubmitted to *Statistics in Medicine*, 8/21/2014

Abstract

Slow recruitment in medical research leads to increased costs and resource utilization, which includes the goodwill contribution of patient volunteers. Careful planning and monitoring of the accrual process can prevent the unnecessary loss of these resources. We propose two hierarchical extensions to the existing Bayesian constant accrual model: the accelerated prior and the hedging prior. The new proposed priors are able to adaptively utilize the researcher's previous experience and current accrual data to produce the estimation of trial completion time. The performance of these models, including prediction precision, coverage probability, and correct decision-making ability, is evaluated using actual studies from our cancer center and simulation. The results showed that a constant accrual model with strongly informative priors works very well when accrual is on target or slightly off, producing smaller mean squared error, high percentage of coverage, and a high number of correct decisions whether or not continue the trial, but it is strongly biased when off target. Flat or weakly informative priors provide protection against an off target prior, but are less efficient when the accrual is on target. The accelerated prior performs similar to a strong prior. The hedging prior performs much like the weak priors when the accrual is extremely off target, but closer to the strong priors when the accrual is on target or only slightly off target. We suggest improvements in these models and propose new models for future research.

Key Words: clinical trials, hedging prior, objective prior, patient accrual, data monitoring committee

1. Introduction

Evaluating and monitoring subject recruitment is important in medical research (Schroen, et al., 2010). Delayed subject recruitment will increase the cost of study and/or can lead researchers to settle for smaller sample sizes than originally proposed. If the proposed sample size is not achieved, the study will have low power and may fail to recognize a treatment effect. Slow recruitment tends to increase resource utilization, which includes the goodwill contribution of patient volunteers. The delayed recruitment may also lead to delay in the adoption of new therapies and slow the advancement of medical progress (Philipson, et al., 2010). For example, a year delay in access to Herceptin would represent a total loss of $8 billion dollars of benefit for breast cancer patients in the United States (Parreco, 2012). Development of reliable and practical tools for accrual prediction is critical in clinical trial studies.

Modeling subject recruitment has been studied for a long time. The available accrual models are reviewed by Barnard (2010) and Zhang (2012a). According to Barnard (2010), the simplest approach is an unconditional model (Carter, 2004). It assumes the accrual rate (e.g., subjects recruited per month) is fixed. The total accrual time is estimated through dividing the planned sample size by the number of subjects they expect to recruit each month. A conditional model allows the accrual rates to vary in any given time depending on other factors that may accelerate or delay the accrual (Carter, Sonne, & Brady, 2005). Compared to the unconditional model, the conditional model matches more closely to real situations. Both unconditional and conditional models fail to consider variations in actual accrual progress, and therefore, their predictions are less accurate. The Poisson models assume that the number of participants recruited within a fixed time follows a Poisson distribution (Carter, Sonne, & Brady, 2005). Anisimov and Fedorov (2007) applied the Poisson model to multi-center clinical trials, assuming

that the rate parameter for the Poisson was different for each center of a multi-center study, and overall the accrual process followed a Poisson-Gamma distribution. Bakhshi, Senn and Phillips (2013) extended the approach with a hierarchical model using historical data from similar trials in the specific center via an empirical Bayesian approach. Some other models have also been proposed, such as the Monte Carlo simulation Markov model by Abbas (2007), time series model by Hadich and Ioannidis (2001), and a model based on nonhomogeneous Poisson process by Zhang and Long (2012b).

Gajewski, Simon and Carlson (2008) developed a Bayesian method for constant accrual based on an exponential waiting model, which is equivalent to Poisson counts over a specified time (Casella & Berger, 2002, p. 100). A unique feature of this model is the incorporation of subjective knowledge about accrual rates through an informative prior distribution (Gajewski, Simon & Carlson, 2008). The strength of the prior distribution is controlled by a parameter P between 0 and 1. If P=1, the prior is given weight equivalent to the proposed sample size of the study. If P=0.5, the prior is given weight equivalent to half the proposed sample size. This means that halfway through the study, the prior and the actual accrual data are given equal weight. If P=0, the prior is effectively ignored (Gajewski, Simon & Carlson, 2008). If the researcher has strong confidence in accrual (for example, he/she has extensive research experience with many similar studies in the same patient setting), the posterior distribution will be weighted heavily towards the prior distribution, especially early in the trial. This avoids an unnecessary alarm when the first few subjects arrive more slowly than expected. On the other hand, if the researcher has a weak prior, early evidence of slow accrual will be given greater weight, encouraging a rapid response to address the slow accrual (Gajewski, Simon & Carlson, 2008).

As discussed above, assessment and specification of control parameter P is critical in simple constant accrual model (Gajewski, Simon & Carlson, 2008). The researchers should constantly monitor/change prior P subjectively during the accrual process, and the assessment of P is totally based on researchers' knowledge and previous experience. There are several limitations regarding this method. One of the most concerns is researchers' over-confidence. What if the researcher provides a strongly informative prior distribution that is substantially off target? This is, unfortunately, quite common. In our experience, researchers are frequently overly optimistic about accrual rates and they are overly confident in their estimates. This experience has been noted by others as well. For example, a study of 78 projects in eight departments of general practice conducted at the Netherlands Institute for Health Services Research and the Centre for Quality of Care Research, found that only 46% recruited the planned number of patients within the planned time frame (van der Wouden, et al., 2007). Other examples of this overconfidence can be found in Williams and Seed (2006), Breau, Carnat, and Gaboury (2006), Jayakaran Saxena, and Yadav (2011), Keen, Pile, and Hill (2005). As we know, mis-spefication of prior by researchers will lead to huge bias in the prediction of accrual, which in turn impairing the management of the trial. Another concern of the existing method is the assessment of P is not always practical. A third party, who lacks sufficient knowledge in the specific trial, may have difficulty to specific P. For example, when the clinical trial is evaluated by the IRB, it will hard for the committee to propose an appropriate subjective prior P to evaluate the trial process. This lack of sufficient message may also introduce bias in P.

To avoid a poor choice of prior mean, a more general and objective approach is to incorporate the historical opinion through power prior (Ibrahim & Chen 2000). The power prior was evaluated and modified by several groups of researchers, which is now well recognized as

the modified power prior (Duan, Ye, & Smith, 2006; Neuenschwander, Branson, & Spiegelhalter 2009). The borrowing strength of historical data is controlled by power parameter. If the power parameter is 0, it borrows on historical information. On the other hand, if the power parameter is 1, it borrows full historical information. Another way to utilize historical data is through commensurate power prior, which is proposed by Hobbs et al. (2011).

In this paper, we propose two adaptive Bayesian priors in monitoring accrual process. The first one is the accelerated prior. The initial prior probability (P) is associated with the proportion of data collected. In the beginning of the trial, when no data are collected, the model will weigh entirely on the investigator's historical experience, with P=1. As data collection progresses, P declines linearly. This puts less weight on the prior and more weight on the data. While all Bayesian models place less weight on the prior as data collection progresses, the accelerated prior speeds up this process to avoid the lingering effects of an off target prior.

The second approach we proposed is the hedging prior. This model treats P as a hyperparameter which follows a uniform distribution. If the prior is off target, the hyperparameter converges to zero as more accrual data arrives. The hedging prior is equivalent to the modified power prior, but the hedging prior is much simpler to implement.

With adaptive priors, P is monitored and determined with the process of the trial. Detailed model descriptions are provided for the accelerated prior and the hedging prior in section 2. Section 3 shows the application and evaluation of the methods using real clinical data. In section 4, the model efficiency and robustness of the new proposed methods and the previous proposed Bayesian model are examined using simulated data under nine different scenarios. Section 5 is the discussion and conclusions.

Model

2.1 Closed form of constant exponential waiting time model

Gajewski, Simon & Carlson (2008) introduced a Bayesian model for constant accrual. Suppose in the original protocol, the investigator planned to recruit *n* subjects in *T* days. Assume that the waiting time (*w*) for each successive patient follows an exponential distribution, $w_i \sim \exp(\theta)$, where θ represents the average accrual time for the *i*th subject. The distribution of the waiting time (*w*) is $f(w|\theta) = \frac{1}{\theta} e^{-w/\theta}$. The prior distribution of θ is assumed to be inverse gamma, $\theta \sim IG(nP, TP)$, where *P* is the investigator's confidence in the original plan, measured on a 0-1 scale. During the trial, m subjects have been collected in T_m ($T_m = \sum_{i=1}^{m} w_i$) time period. Then the posterior distribution for θ is updated to $\theta|\boldsymbol{W} \sim \mathrm{IG}(nP + m, TP + T_m)$, with $\boldsymbol{W} = (w_1, w_2, \ldots w_m)$ (Gajewski, Simon & Carlson, 2008). For fixed n, the waiting time for the rest of the sample size is $\tau = \sum_{i=m+1}^{n} w_i$.

The prediction interval produced by this model provides a rule for stopping a trial with slow accrual. The researcher will specify a decision point, typically 25% or 50% larger than the planned accrual time T, that is, $T_{decision} = 1.25T$ or $T_{decision} = 1.5T$. It represents a delay large enough to threaten the successful completion of a study. If the 95% prediction interval lies entirely above the decision point, then there is sufficient justification to shut down the trial for poor accrual.

We present a new denotation in this paper for the predicted time to recruit fixed number of subjects, as the previous paper used simulation (Gajewski, Simon & Carlson, 2008). The conditional distribution of τ is $\tau \sim \mathrm{G}(m - n, \theta)$, with $f(\tau|\theta) = \frac{1}{\Gamma(n-m)\theta^{n-m}} \tau^{n-m-1} e^{-\tau/\theta}$. Then the predictive distribution of τ can be derived by integration, and we obtain:

$$g(\tau) = \frac{1}{(\mathrm{TP}+T_m)\left(1+\frac{\tau}{\mathrm{TP}+T_m}\right)^{\mathrm{nP+n}}} \left(\frac{\tau}{\mathrm{TP}+T_m}\right)^{\mathrm{n-m-1}} \frac{\Gamma(nP+n)}{\Gamma(\mathrm{nP+m})\Gamma(n-m)} \qquad (1)$$

Let $\phi = \frac{\tau}{\mathrm{TP}+T_m}$, then $\tau = \phi(\mathrm{TP} + T_m)$,

Therefore, $g(\phi)$=$\frac{1}{(1+\phi)^{\mathrm{nP+n}}}\phi^{\mathrm{n-m-1}}\frac{\Gamma(nP+n)}{\Gamma(\mathrm{nP+m})\Gamma(n-m)}$. (2)

It is an inverted beta distribution (Johnson, Kotz & Balakrishnan, 1995), with $\alpha = n - m$, $\beta = nP + m$. When $\beta > 1$, the mean of ϕ is $E(\phi) = \frac{\alpha}{\beta-1} = \frac{\mathrm{n-m}}{\mathrm{nP+m-1}}$. Then the predictive mean of τ, can also be calculated directly as

$$E(\tau) = \frac{\mathrm{n-m}}{\mathrm{nP+m-1}}(\mathrm{TP} + T_m). \quad (3)$$

Similarly, the variance of τ can also be calculated as

$$Var(\tau) = \frac{(\mathrm{n-m})(\mathrm{n+nP-1})}{(\mathrm{nP+m-2})(\mathrm{nP+m-1})^2}(\mathrm{TP} + T_m)^2 \quad (4)$$

The percentile of τ can be obtained by

$$p(\tau) = (\mathrm{TP} + T_m)\frac{p(B)}{1-p(B)}, \quad (5)$$

where p(B) represents the percentile for the beta distribution $beta(n - m, nP + m)$.

To evaluate the sensitivity and stability of the model in accrual prediction, we can also calculate the mean square error. Suppose the true accrual time is T_{truth}, the total recruiting time is $T_p = \tau + T_m$. Then,

$$MSE(T_p) = \frac{(\mathrm{n-m})(\mathrm{n+nP-1})}{(\mathrm{nP+m-2})(\mathrm{nP+m-1})^2}(\mathrm{TP} + T_m)^2 + [\frac{\mathrm{n-m}}{\mathrm{nP+m-1}}(\mathrm{TP} + T_m) + T_m - T_{truth}]^2 . \quad (6)$$

2.2 Hierarchical extensions of the constant accrual model

In the Bayesian accrual model (Gajewski, Simon & Carlson, 2008, 2011), the prior *P* needs to be specified by the researcher. However, in the evaluation by an ethical review committee, it is not practical to do prior specification for each study. In addition, in the monitor of accrual process, it may be too subjective for clinical researchers to choose a single value for the confidence level even if they are familiar with both similar previous trials and the current

study. Therefore, we proposed two adaptive priors for the Bayesian accrual model to make the choice of *P* more objective, as well as to avoid a poor choice of prior mean.

2.2.1 Accelerated Prior

We define the accelerated prior (AP) as

$$P = 1 - \frac{m}{n}. \tag{7}$$

P decreases proportional to the number of subjects recruited. In the beginning of the trial, when there are no subjects recruited, *m* is 0 and *P* is 1. The posterior distribution of θ relies entirely on the prior specification. As more accrual data is collected, the value of *P* will shrink. While all Bayesian models place less weight on the prior distribution as more data is collected, this particular approach accelerates the transition. When m is equal to *n*, *P* will be 0 and the posterior estimation of θ will only be based on collected data.

2.2.2 Hedging Prior

As *P* represents the investigators' confidence in the trial, it actually indicates the similarity of the current trial with historical information. Instead of fixing *P* as a single value, we can set a hierarchical model, specifying the prior distribution for *P* as uniform (0, 1). Therefore, we define the hedging prior (HP) as:

$$\pi(\theta, P|n, T) = \frac{(\mathrm{TP})^{\mathrm{nP}}}{\Gamma(\mathrm{nP})}\left(\frac{1}{\theta}\right)^{nP+1} e^{-\frac{TP}{\theta}} \tag{9}$$

If the prior is off target, the accumulated accrual data will force the distribution of *P* downward, representing a downweighting of the strength of the prior distribution. As more data inconsistent with the off target prior is accumulated, the downweighting should become greater.

The proposed hedging prior is actually a special case of modified power prior. In Bayesian analysis, the modified power prior provides an efficient way to incorporate and down weight historical data (Ibrahim and Chen 2000, Duan 2006, Neuenschwander, Brqanson and

Spiegelhalter 2009). In our exponential waiting time model, the modified power prior can be written as

$$\pi(\theta, P|n, T) = C(P)L(\theta|n, T)^P \pi_0(\theta)\pi(P), \tag{10}$$

Denote the initial prior $\pi_0(\theta) = \frac{1}{\theta}$, which is a special inverse gamma with both shape and scale parameter equaling 0. Assuming the initial prior for P is $\pi(P)$~beta (a,b). When both a and b equal to 1, $\pi(P) = 1$. The researcher's experience and opinion on the accrual is expressed as $L(\theta|n, T) = \theta^{-n}e^{-T/\theta}$. The normalizing coefficient $C(P)$ can be obtained by

$$C(P) = \frac{1}{\int L(\theta|n, T)^P \pi_0(\theta)d\theta}$$

$$= \frac{1}{\int_0^\infty (\theta^{-n}e^{-T/\theta})^P \frac{1}{\theta}d\theta =} = \frac{1}{\int_0^\infty \left(\frac{1}{\theta}\right)^{nP+1} e^{-\frac{TP}{\theta}}d\theta} = \frac{1}{\frac{\Gamma(\mathrm{nP})}{(\mathrm{TP})^{\mathrm{nP}}}} = \frac{(\mathrm{TP})^{\mathrm{nP}}}{\Gamma(\mathrm{nP})}. \tag{11}$$

Finally, the full expression of the prior is

$$\pi(\theta, P|n, T) = C(P)L(\theta|n, T)^P \pi_0(\theta)\pi(P) = \frac{(\mathrm{TP})^{\mathrm{nP}}}{\Gamma(\mathrm{nP})}\left(\frac{1}{\theta}\right)^{nP+1} e^{-\frac{TP}{\theta}}, \tag{12}$$

which is exactly the same as we defined for hedging prior. Thus, the hedging prior is a special case of modified power prior. The hedging prior has a simple and intuitive motivation and can be fit using off the shelf software, such as BUGS.

Using the hedging prior, the posterior distribution of θ and P is

$$\pi(\theta, P|n, T, m, T_m) \propto \frac{(\mathrm{TP})^{\mathrm{nP}}}{\Gamma(\mathrm{nP})}\left(\frac{1}{\theta}\right)^{nP+m+1} e^{-\frac{TP+T_m}{\theta}}, \text{ and} \tag{13}$$

$$\pi(P|n, T, m, T_m) = \int_0^\infty \frac{(\mathrm{TP})^{\mathrm{nP}}}{\Gamma(\mathrm{nP})}\left(\frac{1}{\theta}\right)^{nP+m+1} e^{-\frac{TP+T_m}{\theta}} \mathrm{d}\theta = \frac{(\mathrm{TP})^{\mathrm{nP}}\Gamma(\mathrm{nP+m})}{\Gamma(\mathrm{nP})(TP+T_m)^{nP+m}}, \tag{14}$$

where m is the number of subjects that have been recruited and T_m is the time since trial started, which remains the same as previously defined.

3. Application in clinical trial data

In order to evaluate the performance of the constant accrual model and the two hierarchical extensions, we applied the various models in three clinical studies that were conducted in the University of Kansas Medical Center. Each of these studies was completed prior to the use of our accrual model.

For each of the three studies, the 95% credible intervals of the predicted total accrual time were estimated using three weak priors: a flat prior (P=0), a very weak informative prior (P=0.01), and a weak informative prior (P=0.1). We also considered two strong priors: P=0.5 and P=1. Finally we used the accelerated prior and hedging prior. We evaluated the actual accrual data when 1/8, 1/4, or 1/2 subjects were recruited. The mean square error was calculated according to equation (5).

In addition, we also evaluated whether the methods have the ability to make a correct decision on whether it can alert slow trials. If the accrual is truly not on time, then $T_{truth} > T$. As it would be too strict to use T directly, we define a tolerance coefficient δ, which will produce a decision point $T_{decision} = \delta T$ that can be used for decision making. We also estimated the probability that the predicted total accrual time is less or equal to the cut-off time, which is P($T_p \leq \delta T$). In the current study, the decision point for each trial is set to 1.25T.

3.1 Application Study A: Colorectal Cancer Prevention

Study A investigates "Tailored Touchscreen Computers for Colorectal Cancer Prevention in Urban Core Clinics." In this study, the total proposed sample size is n=460, and total time for accrual is 16 months (487 days). In the real data, the total number of recruited subjects is 471.

Assuming the accrual rate is constant, the corresponding *T* is 561 days. The decision point (701) is set 25% larger than the planned accrual, that is, $T_{decision}$=561×1.25. The total days to recruit 471 subjects is T_{truth}=610 days, which indicates that the trial is slower than planned.

An important feature of this accrual is the lack of constancy over time (Figure 3-1A). The accrual is slightly below target for the first quarter of the time, and even slower as the trial progresses. There is a sudden surge, however, near the end of the trial, not enough to get the trial back on time, but sufficient to cause difficulty for all of the accrual models.

Figure 3-2A shows the performance of each method, assuming 1/8, 1/4, and1/2 of the subjects are recruited. The solid line represents T_{truth}, and the dotted line represents the decision point, 1.25T.

When 1/8 of the data is collected (59 patients), the total accrual time is 88 days. A simple linear extrapolation of this value would indicate that the entire trial would take 704 (88*8) days. All methods provide the point estimates much larger than the planned trial duration of 561 days, with the weaker priors showing more pessimistic estimates than the stronger priors. Both the accelerated prior and the hedging prior behave similar to the stronger priors at this point. The distribution of P for the hedging prior is similar to a uniform distribution, with a mean of 0.446 (versus 0.5 for a uniform) and a 95% interval from 0.027 to 0.971 (versus 0.025 to 0.975 for a uniform). None of the intervals lies entirely above the decision point. The predicted 95% credible interval for the total accrual time covers T_{truth} (610) for each of the models.

As the trial progresses, the accrual starts to slow down. When 1/4 of the data are collected (136 patients), the total accrual time is 201 days. A simple linear extrapolation would indicate a trial duration of 804 days. All of the methods provide very large point estimates, with the weaker priors again producing the most pessimistic estimates. The accelerated prior continues to behave

like a strong prior, but the hedging prior has a point estimate and a prediction interval that more closely approximates a weaker prior. The distribution of P for the hedging prior has shifted markedly, with a mean of 0.08, representing a substantial downweighting of the prior distribution.

Because of a late surge in accrual, all of the methods predict the total accrual time poorly, though the prediction interval for the two strongest priors and the accelerated prior still include T_{truth}. None of the intervals lies entirely above the decision point, indicating (correctly) that the trial should continue.

When 1/2 of the data are collected (236 patients), the total accrual time is 368 days. A simple extrapolation produces an estimated trial duration of 736 days, slightly better than the earlier prediction. All methods are similar, though the point estimates are slightly more pessimistic for the weaker priors and the hedging prior. All of the intervals include the decision point, but they all are completely above T_{truth}. The accelerated prior still behaves much like one of the stronger priors (P=0.5), while the hedging prior behaves much like a weaker prior (P=0.1). The mean for the distribution of P is 0.12, not quite as extreme as earlier, but still a substantial downweighting of the prior distribution. The mean squared error of prediction, which represents overall measure of the performance of each model, is shown on the left of Figure 3-3A. The performance of the strong priors and the accelerated prior seem to be superior to the weak priors and the hedging prior.

Figure 3-3A right shows for each model the estimated probability that the study will finish earlier than the decision point. Notice that the weak priors and the hedging prior are again much more pessimistic than the strong priors and the accelerated prior.

3.2 Application Study B KanQuit 2

KanQuit 2 is a project targeted to treat hospitalized smokers in rural parts of the state via telephone counseling. The proposed total sample size is 576, and *T*=521. Total time to finish recruiting is T_{truth}=1171 days, about twice as long as planned. We set the decision point to 652 (521*1.25). Figure 3-1B shows that the accrual is very slow in the beginning. The accrual rate speeds up slightly when the first 1/3 of the subjects are recruited.

Figure 3-2B shows the coverage of predicted total accrual time using each method. Although there is substantial variation in the location and width of the individual prediction intervals, it is worth noting that every single model would have recommended early termination of the study even only 1/8 of the data was collected.

With 1/8 of the data is collected (72 patients in 244 days), the simple linear extrapolation would estimate the total completion time to be 1,952 days, more than triple of the planned accrual rate. The weak priors are close to this estimate. The strong priors are less pessimistic, but even these priors, which weight the data much less strongly than the off target prior, still are pessimistic enough to recommend early termination for slow accrual. The accelerated prior behaves much like the strong prior. The hedging prior behaves much like a weak prior. The mean of the distribution of *P* in the hedging prior is 0.004, representing an almost total rejection of the off target prior. The results for all models at 1/4 and 1/2 of the subjects are similar.

Notice that the stronger priors are actually better at including T_{truth}. The downward pull of the off target prior prevents these models from overreacting to the temporary slow accrual at the start of the study.

Figure 3-3B left shows the mean squared error of prediction. The pattern is similar to Figure 3-3A. Figure 3-3B right shows that for this extreme study, all the models estimate the

probability that the study will finish earlier than the decision point to be zero, even with only 1/8 of the accrual data available.

3.3 Application Study C KIS III study

This study evaluates the efficacy of sustained release bupropion in combination with health education (HE) counseling for smoking cessation among urban African American light smokers. The study is a two-arm, double-blinded, placebo-controlled design.

The total proposed sample size is n=540. The proposed total time for accrual is 24 months, with T=730 days. The decision point is 912 (1.25*730). The real time for accrual is T_{truth}=670. There is also some minor variation in accrual rates. The study was slightly slow in the very beginning. Then it sped up and finished earlier than expected. Overall, this study represents a rather small change (less than 10%), and it would be fair to characterize the prior distribution as just very slightly off target.

All the methods produced similar point estimates, which is expected for a prior distribution that is only slightly off target. None of the methods would suggest early termination of the trial for poor accrual (Figure 3-2C). Note that the prediction intervals are narrower for the strong priors. The accelerated prior behaves much like the strong prior, but the hedging prior also behaves like a strong prior. The distribution of P is similar to a uniform distribution—the mean is 0.446, 0.507, and 0.496 respectively. All of the prediction intervals cover T_{truth}.

Figure 3-3C right shows the estimated probability that the study will finish earlier than the decision point by each model. Since this study was actually accruing patients slightly faster than planned, each of the models estimates this probability effectively at 100%.

4. Simulation Studies

Motivated by the application of the various methods in real clinical studies, a simulation study was designed to evaluate the robustness and performance of proposed accelerated prior and hedging prior, and compare them to the constant accrual model.

We assume that a researcher proposed to recruit 300 subjects in 3 years. Thus *n* equals 300, and *T* is 1095 days. If the accrual is on target, the theoretical waiting time θ_0 is 3.65 days/patient recruited. We also assume that the waiting time for slow accrual θ_1 is 7.31 days/patient (off target), and for fast accrual θ_2 is 1.83 days/patient (off target). The waiting times for all 300 subjects are simulated based on the assumption that the waiting time for each subject is independently distributed as exponential (θ), or piecewise exponential (θ) to reflect a non-homogenous process.

Based on the different combinations of θ_0 (on target accrual), θ_1 (off target, slow accrual), θ_2 (off target, fast accrual). We designed nine studies to mimic the situations commonly encountered in real clinical trial studies. Study 1 represents an unbiased situation, where the whole accrual process is on target. The waiting time for each subject is exponential with $\theta_i = \theta_0$ (i is from 1 to 300). Study 2 represents a slow accrual, in which the waiting time is two times as long as planned, during the whole study. Study 3 is a fast accrual and the accrual time is only half of the proposed time. The other six studies resemble a step wise accrual process. Study 4 shows a situation where the accrual is slow for the first tenth of the subjects, $\theta_i = \theta_1$ (i is from 1 to 30). The recruitment then goes back to normal for the rest of the subjects $\theta_i = \theta_0$ (i is from 31 to 300). Study 5 mimics a trial that is slow at both the first and last tenth of the subjects. Study 6 is similar to study 2, with the exception that the slow accrual period is longer, such that the first quarter of the subjects is recruited slowly. Study 7 is parallel with study 3, with slow accrual for both first and last quarter of subjects. Study 8 represents a trial that starts with slow accrual at the

first quarter, goes back to normal on the second quarter, and then catches up quickly to the schedule during the last half of accrual, with $\theta_i = \theta_2$ (i is from 151 to 300). Study 9 is an example of a trial that is on time for the first half subjects, but lags for the last half. The theoretical settings of the accrual process for each study are shown in Table 1 and displayed in Figure 3-4. Each simulation study is repeated for 1000 times. For each simulation, T_{truth} is the sum of the waiting times, calculated as $T_{truth} = \sum_{i=1}^{300} w_i$. The 2.5%, 50%, and 97.5% quantiles of T_{truth} are calculated from the 1000 simulated samples. In Figure 3-4, for comparison purpose, the dotted line represents the theoretical accrual rate if the recruitment is unbiased or on target. The vertical dash line shows the theoretical time (*T*) for which the trial should be finished.

For each set of simulated data, we assume the first 1/8, 1/4, 1/2, 3/4 of the data are known, with *m*=37, 75, 150, 225, respectively. The total predicted accrual time is estimated using conjugate inverse gamma priors and are specified as *P*=0, *P*=0.01, *P*=0.1, *P*=0.5 and *P*=1, as well as the proposed accelerated prior and hedging prior. For each method in a particular study, the mean squared error of total predicted time is calculated as $= \frac{1}{1000}\sum_{j=1}^{1000} MSE_j$, and the average of the relative bias for each study is calculated as $RBIAS = \frac{1}{1000}\sum_{j=1}^{1000} \frac{\hat{T} - Ttruth_j}{Ttruth_j} \times 100\%$. The summary of MSE and $RBIAS$ are shown in Figure 3-5 and Figure 3-6, respectively. We did comparison and summarized the results for all the seven methods. As we found that the results between P=0 and P=0.01are very similar, we exclude P=0.01 from Figure 3-5 and 3-6 for the purpose of better display.

For the first simulation study, when the whole accrual process is on target, it is not surprising that the informative prior with *P*=1 has the smallest MSE. The accelerated prior behaves very similarly to *P*=0.5, which is almost as good as *P*=1, when 1/8, 1/4, 1/2, 3/4 of the data are known. The hedging prior has a slightly larger MSE, although much better than when P

is very small, P=0.1 or P=0.01, or with non-informative prior (P=0). However, for a biased accrual (study 2 and 3), either slow or fast, informative prior P=1 has the largest MSE. P=0, P=0.01, P=0.1 and hedging prior have smaller MSE. The accelerated prior has an MSE comparable to the strong priors in the beginning of the trial, but have smaller MSE when $m = \frac{3}{4}n$.

When the first 10% of subjects are slow in accrual (simulation study 4), the results of MSE are similar to study 1. The strong priors produce the smallest MSE. The weak priors have much larger MSE values. The MSEs of accelerated prior are comparable to the strong priors. The hedging prior suffers from the same problems as the weak priors, giving too much weight to the early slow accrual data. The results from study 5 (slow first and last 1/10) are very similar to study 4. The remaining studies 6, 7, 8 show similar results. If the initial early accrual is slow, the weak priors and the hedging prior give too much weight to this data. They can recover at times when there is sufficient accrual data beyond the initial slow accrual period (that is, at m= 150). The advantage of the strong priors is that they are not unduly swayed by the early slow accrual. The methods all behavior similarly for study 9.

In each of the simulation studies, we not only estimated the total accrual time T using all different methods, but also calculated the 95% credible interval $T_{0.025}$ and $T_{0.975}$. The correct coverage is defined as $CCj = I(Ttruthj \geq T_{0.025}) + I(Ttruthj \leq T_{0.975})$. The total percentage of correct coverage is $CC = \frac{1}{1000}\sum_{j=1}^{1000} CC_j \times 100\%$. The summary of the correct coverage is shown in Table 2. For unbiased accrual, the coverage probabilities of all methods are around or higher than 95%. When the accrual is either slow or fast, P0, P0.01 and hedging prior provide higher coverage all the time, where informative priors (P1, P0.5, AP) fail to cover T_{truth}.

In contrast, the weak priors perform poorly if there are short slow accrual periods, as in study 4 and 5. The hedging prior has poor coverage probability early, but seems to recover with a larger amount of accrual data beyond the slow initial period. Study 6, with a longer period of slow accrual, appears to cause problems for all of the models. Study 7, with slow accrual both during the first quarter and last quarter of the study, no model is consistently accurate. All of the methods have low coverage for study 8 and 9.

More important, perhaps, than the percentage of coverage is the ability of the models to make a correct decision on the continuation of the trial. Similar as application studies in the previous section, we defined a cut-off time δT that can be used for decision making. If $\hat{T}_{0.025} > \delta T$, we should stop the trial and the decision is NOGO. If $\hat{T}_{0.025} \leq \delta T$, the decision is GO. For one simulation iteration, the correct decision can be calculated as $CDj = I(\hat{T}_{0.025} > \delta T)$ $I(Ttruth_j > T)$+ $I(\hat{T}_{0.025} \leq \delta T)$ $I(Ttruth_j \leq T)$. The percentage of the correct decisions is $CD = \frac{1}{1000}\sum_{j=1}^{1000} CD_j \times 100\%$. In this simulation study, δ is also set up as 1.25.

For the unbiased trial, all models correctly recommend continuation of the study. For the slow accrual trial, all models also correctly recommend stopping the trial, but the recommendations to stop are more likely to come early for the weak priors and the hedging prior. For the fast trial, all models correctly recommend continuation.

The time varying models provide a much more difficult challenge. Study 4 has slow accrual during the first tenth of the trial, but the trial still finishes earlier than the decision point. All of the models correctly recommend continuation of the trial when $m = \frac{1}{2}n$, but the weak priors and the hedging prior will too frequently recommend termination of the study when $m = \frac{1}{8}n$ and $m = \frac{1}{4}n$. For study 5, 25% of the decisions are incorrect with every model, but the two

weakest priors (P=0 and P=0.01) perform even worse at $m = \frac{1}{8}n$.In study 6, the weak priors perform slightly better at $m = \frac{1}{8}n$ and the strong priors perform slightly better at $m = \frac{1}{4}n$ and $m = \frac{1}{2}n$, but the performance of all the models is mediocre. Performance is equally mediocre for Study 7, with slow early accrual during the first quarter and last quarter of the trial. The strong priors perform especially poorly here. In contrast, the weak priors perform especially poorly for Study 8 especially when $m < \frac{1}{2}n$. The probability of correction decision in study 9 is zero for all methods during the first half of the trail, slightly better $m = \frac{3}{4}n$.

5. Discussion

Monitoring the accrual process in clinical trials is critical, as slow recruitment results in increasing costs, utilizing more resources, and wasting the goodwill contribution of patient volunteers.

From clinical studies, we find that that all of the models perform well when the prior is far off target. Even when they do not predict the trial completion time accurately, they all recognize early that the trial is off schedule. The performance of the strong priors is a bit surprising, but perhaps some of this is an accidental result of the surge in accrual later on the studies. The accelerated prior does not appear to behave much differently than a strong prior in any of the studies. In contrast, the hedging prior did seem to adapt its behavior somewhat, behaving more like a weak prior when the accrual was substantially off target, but like a strong prior when the accrual was only slightly off target. One important lesson, though, is that variations in accrual rates can complicate the evaluation of these models.

When there is variation in accrual rates, strong priors appear to perform better. This result is somewhat surprising. Placing a high weight on an off target informative prior would seem to

be a bad idea. Perhaps, though, the weak priors are placing too much weight on the slow early accrual which is not going to last throughout the trial. Overall, strong priors are likely to overreact to a bit of slow accrual data early in the study. But every prior, weak or strong, has difficulty with at least some of the time varying scenarios.

The simulation studies show that the informative prior, especially strong informative prior, works better when the accrual is on target, or slightly off from the target, producing smaller MSE, higher percentage of coverage, and more correct decisions. However, they are terrible when the accrual is entirely off from the target. Flat prior and weak informative prior are less efficient when the accrual is on target, but work well for very slow accrual.

Performance of the accelerated prior is typically similar to the strong priors in the early period of trial. However, it is superior to strong priors when more data are available for off target accruals. The accelerated prior was designed to transition rapidly from a strong to a weak prior when more accrual data is available. The accelerated prior is equivalent to P=0.5 when half the data is collected, and this is still a strong prior distribution. It might be worth examining an even stronger degree of acceleration, such as a cubed acceleration, $P = \left(1 - \frac{m}{n}\right)^3$, which would be equivalent to P=0.125 when half of the data is collected. Another alternate form of P would be set $P = 1 - \frac{2m}{n}$ for m≤n/2, and P=0 for m>n/2. The degree of acceleration will be stronger than the accelerated prior we proposed. It has intuitive appeal, as in most trials, the halfway point may be considered a reasonable time to rely on the data that have been observed.

The performance of the hedging prior is difficult to characterize well, but it appears to perform better than strong priors when the prior is extremely off target, but also superior to flat priors when the prior is on target or only slightly off target. We have shown that the hedging prior is equivalent to the modified power prior, but it is easier to fit. Instead of monitoring the

accrual process and assessment of the prior P subjectively, the hedging prior evaluates the similarity of real accrual process and historical data simultaneously, and adjusts the weights of the prior thereafter. It could be a useful compromise between the completely data driven flat prior and the reliance on investigator's opinion that a strong prior offers. Hedging prior is a special case of modified power prior. However, it avoids an extra "missing integral step". More importantly, hedging prior preserves the prior interpretation found in original constant accrual model.

As the non-constant accrual rate changes over time. If we know for sure that the accrual can be divided into a certain number of stages and an appropriate prior can be selected for each specific stage. Then the strong prior will perform well in accrual prediction. However, there are many factors that affect the accrual process. The change time point of the accrual and its direction (slow or fast) is unpredictable, such as simulation study 8 and 9. It is hard to know what the accrual will be in certain stage and it is impossible choose the appropriate informative prior for each stage. Actually, there are examples of Bayesian regression models that could be adapted to this problem. We plan to consider some of these models in a future publication.

Overall, there is no obvious choices among the various models proposed here. There are trade-offs between the greater precision provided by the strong priors and the ability of the weak priors to recognize more rapidly problems with slow accrual.

Acknowledgments

This work was supported by an NIH grant from NCI awarded to The University of Kansas Cancer Center (The KUCC) #1 P30 CA168524. We thank Dr. Allen Greiner, Dr. Edward Ellerbeck, and Dr. Lisa Cox for providing their patient accrual data in their research projects, which were funded by NIH (R01CA123245, R01CA101963, R01CA091912 and R01CA091912-09S1)

Appendix: Mathematical formulas for deriving closed form of constant exponential waiting time model

Assuming that the waiting time (w) for each successive patient follows an exponential distribution, $w_i \sim \exp(\theta)$, where θ represents the average accrual time for the ith subject. The distribution of the waiting time (w) is $f(w|\theta) = \frac{1}{\theta} e^{-w/\theta}$. The prior distribution of θ is assumed to be invers gamma, $\theta \sim IG(nP, TP)$, where P is the investigator's confidence on the original plan, measured on a 0-1 scale. Supposedly during a trial, m subjects have been collected in T_m ($T_m = \sum_{i=1}^{m} w_i$) time period. Then the posterior distribution for θ is updated to $\theta|w \sim \mathrm{IG}(nP + m, TP + T_m)$, which can also be written as:

$$f(\theta|m, T_m) = \frac{(TP+T_m)^{nP+m}}{\Gamma(nP+m)} \theta^{-(nP+m+1)} e^{-\frac{TP+T_m}{\theta}}. \quad (1)$$

For fixed n, the waiting time for the rest of the sample size is $\tau = \sum_{i=m+1}^{n} w_i$.

We present a new denotation in this paper for the predicted time to finish for fixed sample size, as the previous paper used simulation. The distribution of τ is $\tau \sim \mathrm{G}(m-n, \theta)$, with $f(\tau|\theta) = \frac{1}{\Gamma(n-m)\theta^{n-m}} \tau^{n-m-1} e^{-\tau/\theta}$. Then the predictive distribution of τ can be derived as following,

$$g(\tau) = \int_0^{\infty} \frac{(\mathrm{TP}+T_m)^{\mathrm{nP+m}}}{\Gamma(\mathrm{nP+m})} \theta^{-(nP+m+1)} e^{-(TP+T_m)/\theta} \frac{1}{\Gamma(n-m)\theta^{n-m}} \tau^{n-m-1} e^{-\tau/\theta} d\theta \quad (2)$$

$$= \frac{(\mathrm{TP}+T_m)^{\mathrm{nP+m}} \tau^{n-m-1}}{\Gamma(\mathrm{nP+m})\Gamma(n-m)} \frac{\Gamma(nP+n)}{(\mathrm{TP}+T_m+\tau)^{\mathrm{nP+n}}}$$

$$= \frac{1}{(\mathrm{TP}+T_m)\left(1+\frac{\tau}{\mathrm{TP}+T_m}\right)^{\mathrm{nP+n}}} \left(\frac{\tau}{\mathrm{TP}+T_m}\right)^{\mathrm{n-m-1}} \frac{\Gamma(nP+n)}{\Gamma(\mathrm{nP+m})\Gamma(n-m)}. \quad (3)$$

Let $\phi = \frac{\tau}{\mathrm{TP}+T_m}$, then $\tau = \phi(\mathrm{TP} + T_m)$,

Therefore, $g(\phi) = \frac{1}{(\mathrm{TP}+T_m)\left(1+\frac{\tau}{\mathrm{TP}+T_m}\right)^{\mathrm{nP+n}}} \left(\frac{\tau}{\mathrm{TP}+T_m}\right)^{\mathrm{n-m-1}} \frac{\Gamma(nP+n)}{\Gamma(\mathrm{nP+m})\Gamma(n-m)} |\mathrm{TP} + T_m|$,

which can be simplified as $g(\phi)=\frac{1}{(1+\phi)^{\mathrm{nP+n}}}\phi^{\mathrm{n-m-1}}\frac{\Gamma(nP+n)}{\Gamma(\mathrm{nP+m})\Gamma(n-m)}$. (4)

It is an inverted beta distribution, with $\alpha = n - m$, $\beta = nP + m$. When $\beta > 1$, the mean of ϕ is $E(\phi) = \frac{\alpha}{\beta-1} = \frac{\mathrm{n-m}}{\mathrm{nP+m-1}}$. Then the predictive mean of τ, can also be calculated directly as

$$E(\tau) = E(\phi)(\mathrm{TP} + T_m) = \frac{\mathrm{n-m}}{\mathrm{nP+m-1}}(\mathrm{TP} + T_m). \quad (5)$$

Similarly, the variance of τ can also be calculated as

$$Var(\tau) = Var\big(\phi(\mathrm{TP} + T_m)\big) = \frac{(\mathrm{n-m})(\mathrm{n+nP-1})}{(\mathrm{nP+m-2})(\mathrm{nP+m-1})^2}(\mathrm{TP} + T_m)^2 \quad (6)$$

The percentile of τ can be obtained by

$$p(\tau) = (\mathrm{TP} + T_m)p(\phi) = (\mathrm{TP} + T_m)\frac{p(B)}{1-p(B)}, \quad (7)$$

where p(B) represents the percentile for the beta distribution $beta(n - m, nP + m)$.

To evaluate the sensitivity and stability of the model in accrual prediction, we can also calculate the mean square error. Suppose the true accrual time is T_{truth}, the predicted total recruiting time is $T_p = \tau + T_m$. Then,

$$MSE(T_p) = Var(T_p) + [E\big(T_p - T_{truth}\big)]^2$$

$$= \frac{(\mathrm{n-m})(\mathrm{n+nP-1})}{(\mathrm{nP+m-2})(\mathrm{nP+m-1})^2}(\mathrm{TP} + T_m)^2 + [\frac{\mathrm{n-m}}{\mathrm{nP+m-1}}(\mathrm{TP} + T_m) + T_m - T_{truth}]^2. \quad (8)$$

The probability that the total recruiting time is larger than critical cut-off can be calculated as the following:

$$P\big(T_p \le \delta T\big) = P(\tau + T_m \le \delta T) = P(\tau \le \delta T - T_m) = P\left(\frac{\tau}{\mathrm{TP} + T_m} \le \frac{\delta T - T_m}{\mathrm{TP} + T_m}\right)$$

$$= P\left(\phi \le \frac{\delta T - T_m}{\mathrm{TP}+T_m}\right). \quad (9)$$

Table 1. The design of the eight simulation studies, including short description of the study, parameter setup, and 2.5%, 50, and 97.5% quantile of Ttruth with 1000 iterations.

Simulation Study	Parameter setup	Ttruth 2.5%	Ttruth 50%	Ttruth 97.5%
1 "Unbiased"	$\theta_i = \theta_0$ for i=1 to 300	973.2	1094.9	1213.2
2 "Slow"	$\theta_i = \theta_1$ for i=1 to 300	1946.5	2189.8	2426.5
3 "Fast"	$\theta_i = \theta_2$ for i=1 to 300	486.6	547.5	606.6
4 "Slow early 1/10"	$\theta_i = \theta_1$ for i=1 to 30	1072.6	1202.6	1336.9
	$\theta_i = \theta_0$ for i=31 to 300			
5 "Slow early and last 1/10"	$\theta_i = \theta_1$ for i=1 to 30	1171.3	1312.7	1469.9
	$\theta_i = \theta_0$ for i=31 to 270			
6 "Slow early ¼"	$\theta_i = \theta_1$ for i=1 to 35	1211.4	1367.8	1523.5
	$\theta_i = \theta_0$ for i=76 to 300			
7 "Slow early and last ¼"	$\theta_i = \theta_1$ for i=1 to 75	1455.0	1642.0	1831.7
	$\theta_i = \theta_0$ for i=76 to 225			
	$\theta_i = \theta_1$ for i=226 to 300			
8 " Slow early ¼ and fast last ½	$\theta_i = \theta_1$ for i=1 to 75	962.0	1091.3	1237.6
	$\theta_i = \theta_0$ for i=76 to 150			
	$\theta_i = \theta_2$ for i=151 to 300			
9 " On time first ½ and slow last ½	$\theta_i = \theta_0$ for i=1 to 150	1451.8	1641.3	1827.9
	$\theta_i = \theta_1$ for i=150 to 300			

Table 3-1. The design of the eight simulation studies, including short description of the study, parameter setup, and 2.5%, 50, and 97.5% quantile of Ttruth with 1000 iterations.

Simulation Study	Parameter setup	Ttruth 2.5%	Ttruth 50%	Ttruth 97.5%
1 "Unbiased"	$\theta_i = \theta_0$ for i=1 to 300	973.2	1094.9	1213.2
2 "Slow"	$\theta_i = \theta_1$ for i=1 to 300	1946.5	2189.8	2426.5
3 "Fast"	$\theta_i = \theta_2$ for i=1 to 300	486.6	547.5	606.6
4 "Slow early 1/10"	$\theta_i = \theta_1$ for i=1 to 30	1072.6	1202.6	1336.9
	$\theta_i = \theta_0$ for i=31 to 300			
5 "Slow early and last 1/10"	$\theta_i = \theta_1$ for i=1 to 30	1171.3	1312.7	1469.9
	$\theta_i = \theta_0$ for i=31 to 270			
6 "Slow early ¼"	$\theta_i = \theta_1$ for i=1 to 35	1211.4	1367.8	1523.5
	$\theta_i = \theta_0$ for i=76 to 300			
7 "Slow early and last ¼"	$\theta_i = \theta_1$ for i=1 to 75	1455.0	1642.0	1831.7
	$\theta_i = \theta_0$ for i=76 to 225			
	$\theta_i = \theta_1$ for i=226 to 300			
8 " Slow early ¼ and fast last ½	$\theta_i = \theta_1$ for i=1 to 75	962.0	1091.3	1237.6
	$\theta_i = \theta_0$ for i=76 to 150			
	$\theta_i = \theta_2$ for i=151 to 300			
9 " On time first ½ and slow last ½	$\theta_i = \theta_0$ for i=1 to 150	1451.8	1641.3	1827.9
	$\theta_i = \theta_1$ for i=150 to 300			

Table 3-2. Summary of the percentage of coverage of T_{truth} and the percentage of correct decisions to either continue or stop the trial (shaded) using the various methods when recruited the first 1/8, 1/4 or 1/2 of the subjects in the eight simulation studies for 1000 iterations.

	Method	Unbiased	Slow	Fast	Slow early 1/10	Slow early and last 1/10	Slow early ¼	Slow early and last ¼	Slow early ¼ and fast last ½	On time first ½ and lags last ½
		1	2	3	4	5	6	7	8	9
1/8	P0	96	95	95	8	18	10	50	0	31
	P0.01	97	94	95	9	21	13	56	0	27
	P0.1	99	60	12	19	48	40	92	1	6
	P0.5	100	0	0	65	94	95	54	5	0
	P1	99	0	0	85	97	97	10	12	0
	AP	99	0	0	82	98	98	15	10	0
	HP	100	85	78	35	59	45	74	3	1
1/4	P0	95	96	95	26	60	0	14	0	3
	P0.01	95	95	95	27	63	0	16	0	2
	P0.1	97	77	40	40	80	1	45	0	1
	P0.5	98	1	0	75	97	13	96	0	0
	P1	98	0	0	88	94	44	76	0	0
	AP	98	0	0	83	96	30	90	0	0
	HP	98	92	91	59	87	2	32	0	1
1/2	P0	95	95	95	61	95	6	95	0	0
	P0.01	95	94	95	62	96	8	94	0	0
	P0.1	96	88	74	70	96	12	91	0	0
	P0.5	97	19	1	85	87	40	55	0	0
	P1	97	1	0	90	78	65	22	0	0
	AP	97	19	1	85	87	40	55	0	0
	HP	96	94	94	81	91	18	89	0	0
3/4	P0	96	95	95	85	65	46	13	0	13
	P0.01	96	95	95	85	64	47	12	0	12
	P0.1	96	93	89	87	59	55	8	0	8
	P0.5	96	58	24	91	45	73	1	0	2
	P1	96	20	4	92	36	84	1	0	0
	AP	96	82	64	89	53	64	4	0	5
	HP	96	95	95	91	50	62	7	0	8

Table 3- 3 Summary of the percentage of correct decisions to either continue or stop the trial (shaded) using the various methods when recruited the first 1/8, 1/4 or 1/2 of the subjects in the eight simulation studies for 1000 iterations.

	Method	Unbiased	Slow	Fast	Slow early 1/10	Slow early and last 1/10	Slow early ¼	Slow early and last ¼	Slow early ¼ and fast last ½	On time first ½ and lags last ½
		1	2	3	4	5	6	7	8	9
1/8	P0	100	88	100	31	50	58	85	15	0
	P0.01	100	86	100	35	53	60	82	18	0
	P0.1	100	61	100	61	69	67	58	40	0
	P0.5	100	8	100	97	76	58	8	94	0
	P1	100	1	100	99	75	54	2	100	0
	AP	100	2	100	99	75	54	2	99	0
	HP	100	57	100	67	70	67	55	44	0
1/4	P0	100	99	100	73	74	49	99	2	0
	P0.01	100	99	100	76	75	49	98	2	0
	P0.1	100	98	100	88	78	51	96	4	0
	P0.5	100	79	100	98	76	67	78	24	0
	P1	100	56	100	99	75	76	53	48	0
	AP	100	65	100	99	75	73	64	38	0
	HP	100	95	100	91	78	52	95	5	0
1/2	P0	100	100	100	97	78	67	80	20	0
	P0.01	100	100	100	97	77	67	79	21	0
	P0.1	100	100	100	98	76	72	74	27	0
	P0.5	100	100	100	99	75	81	55	47	0
	P1	100	100	100	99	75	82	41	61	0
	AP	100	100	100	99	75	81	55	47	0
	HP	100	100	100	98	76	76	69	33	0
3/4	P0	100	100	100	99	76	88	50	97	50
	P0.01	100	100	100	99	76	88	50	97	49
	P0.1	100	100	100	99	76	88	47	98	47
	P0.5	100	100	100	99	75	87	39	98	41
	P1	100	100	100	99	75	85	34	99	37
	AP	100	100	100	99	75	88	43	98	44
	HP	100	100	100	99	75	88	44	98	45

Figure 3-1. The accumulated accrual for application studies, A (TSCCP), B (KanQuit2), C (KISIII). The solid line is the real accrual. The dotted line is proposed reference, and the vertical dash line shows the proposed T.

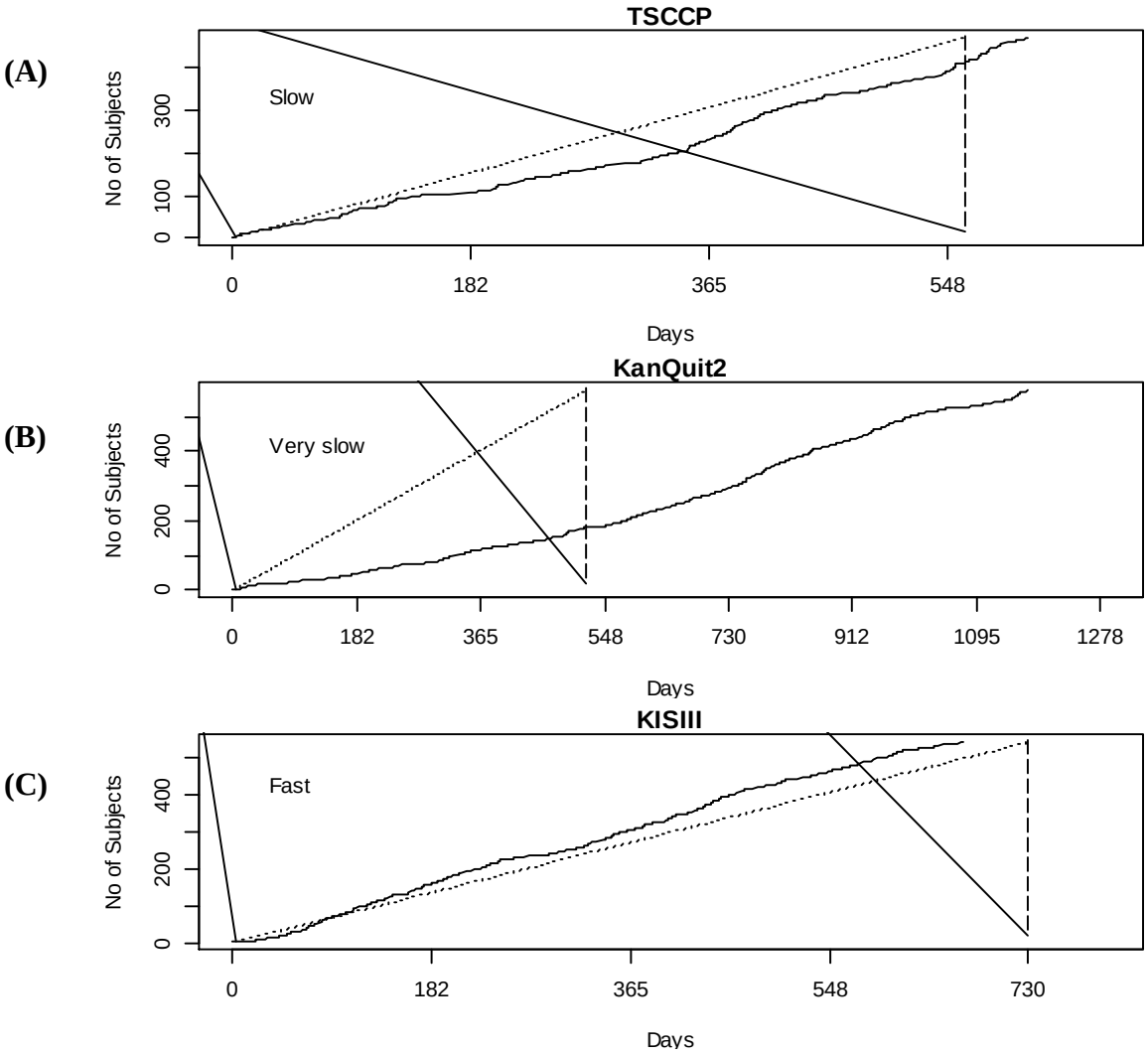
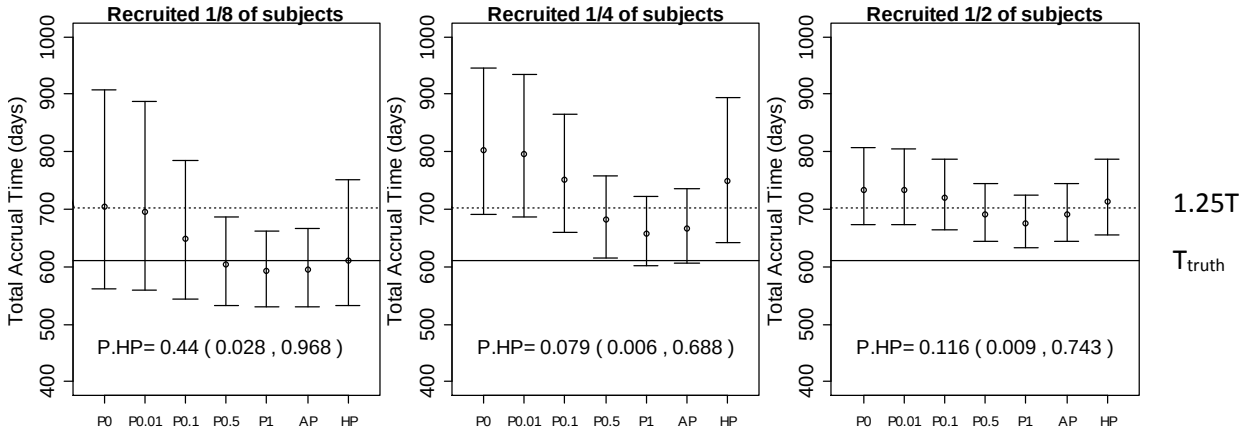
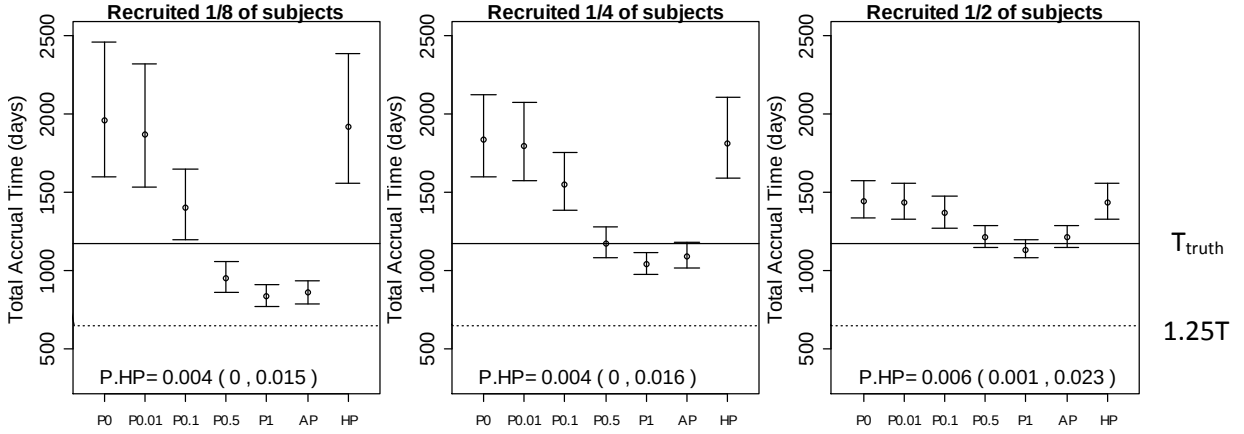


Figure 3-2. The prediction of total accrual time for each of the studies A (cancer), B (KanQuit2), C (KISIII) using various methods assuming only 1/8, 1/4, and 1/2 of the subjects recruited. The solid line shows the true accrual time, and the dotted line shows the true decision line that should stop the trial.

(A)



(B)



(C)

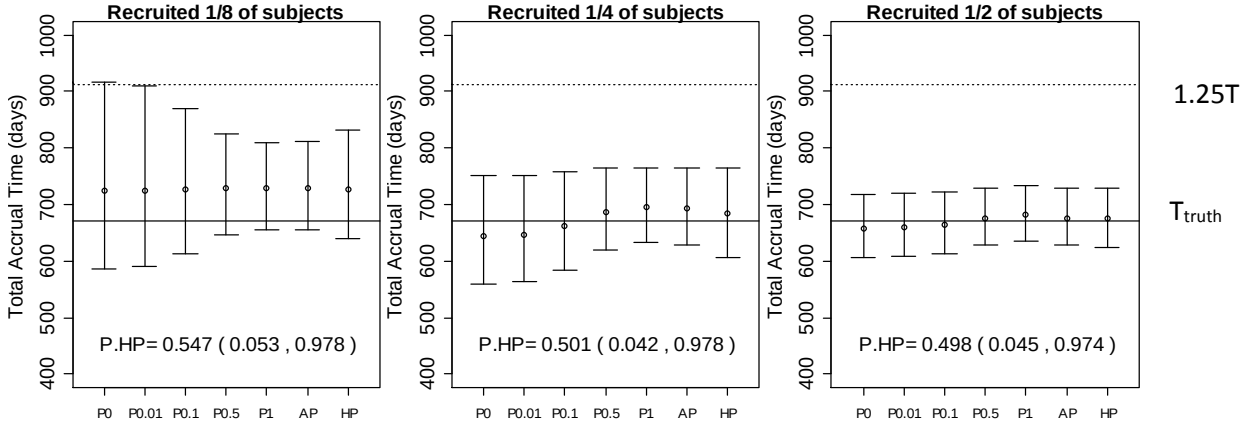


Figure 3-3. The mean squared error (left) and the probability that the predicted accrual time is less or equal to the cut-off time (right) for each of the application studies: A (TSCCP), B (KanQuit2), C (KISIII), study using various methods assuming only 1/8, 1/4, and 1/2 of subjects recruited. (The labels in the figure of Probablity of Stop Trial for the KanQuit 2 and KIS III study are overlapped is because the results are exactly the same).

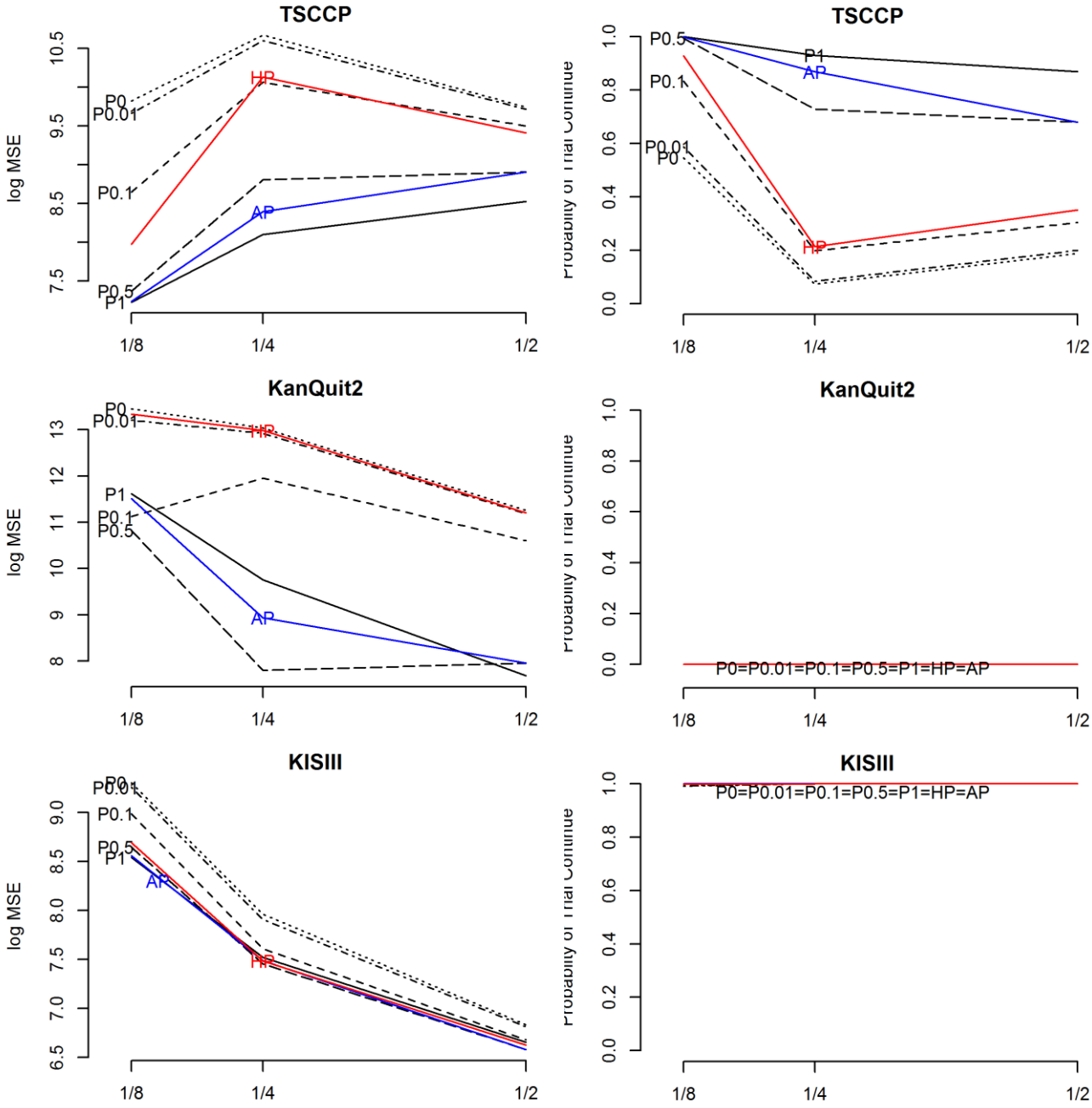


Figure 3-4. The graphical display of the theoretical accrual process of the eight studies. The solid line is the designed simulation studies, dotted line is the reference if the trial is on target, and the vertical dash line shows the proposed T.

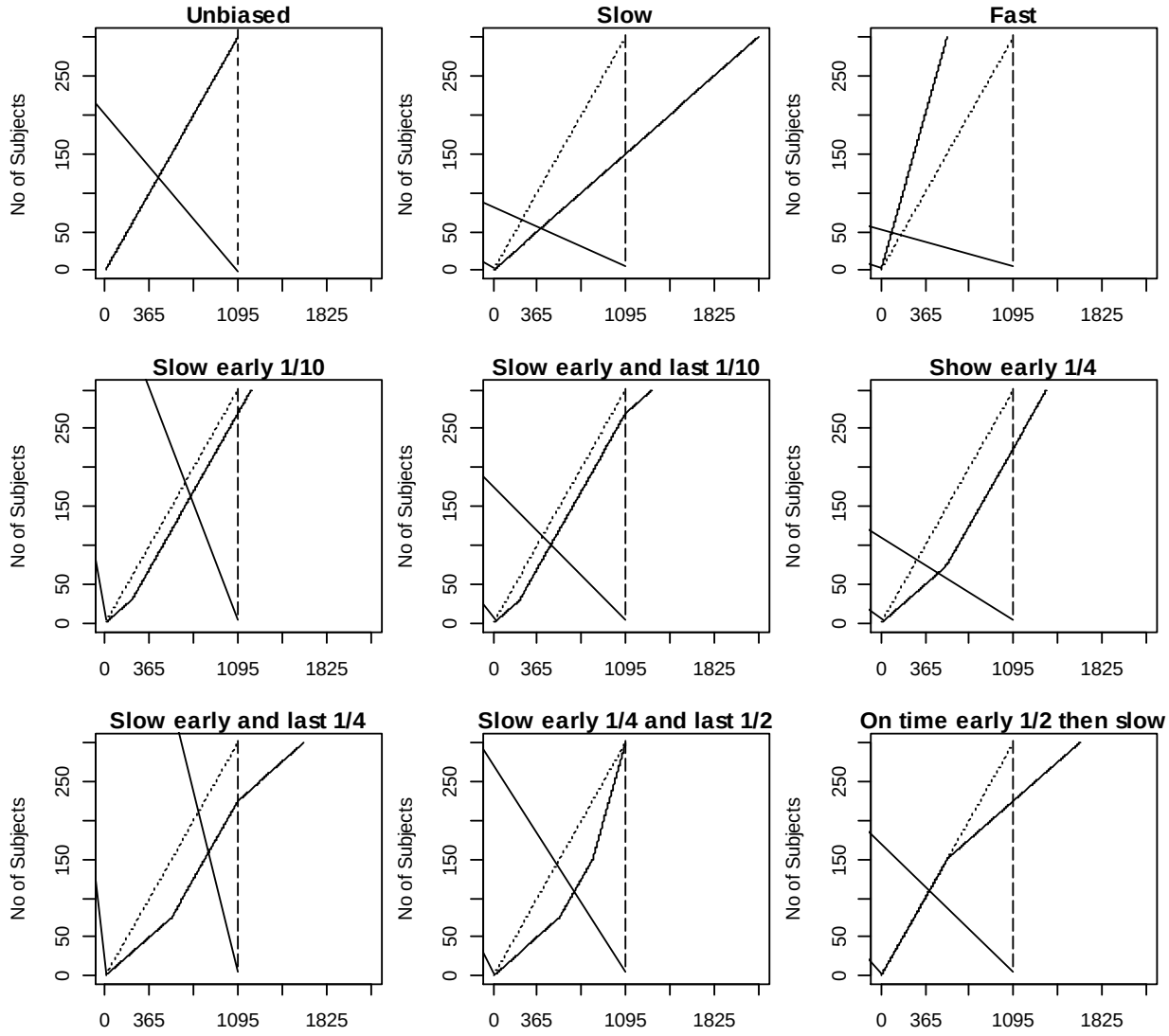


Figure 3-5. The MSE on log scale of each simulation study using the various methods (P0, P0.1, P0.5, AP, HP) when recruited the first 1/8, 1/4, 1/2 or 3/4 of the subjects. As the results for study 9 On time early 1/2 then slow are too close, their labels are overlapped.

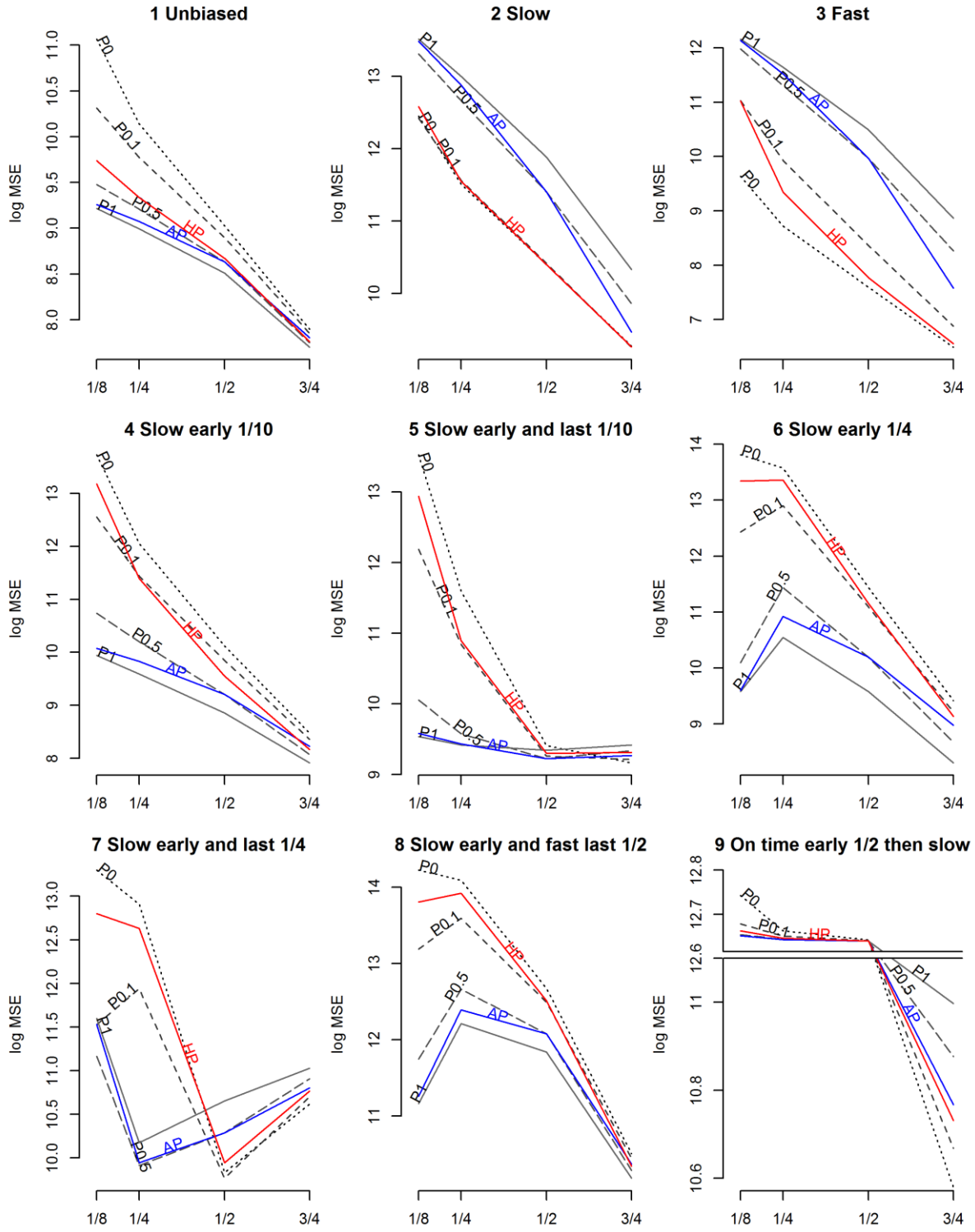
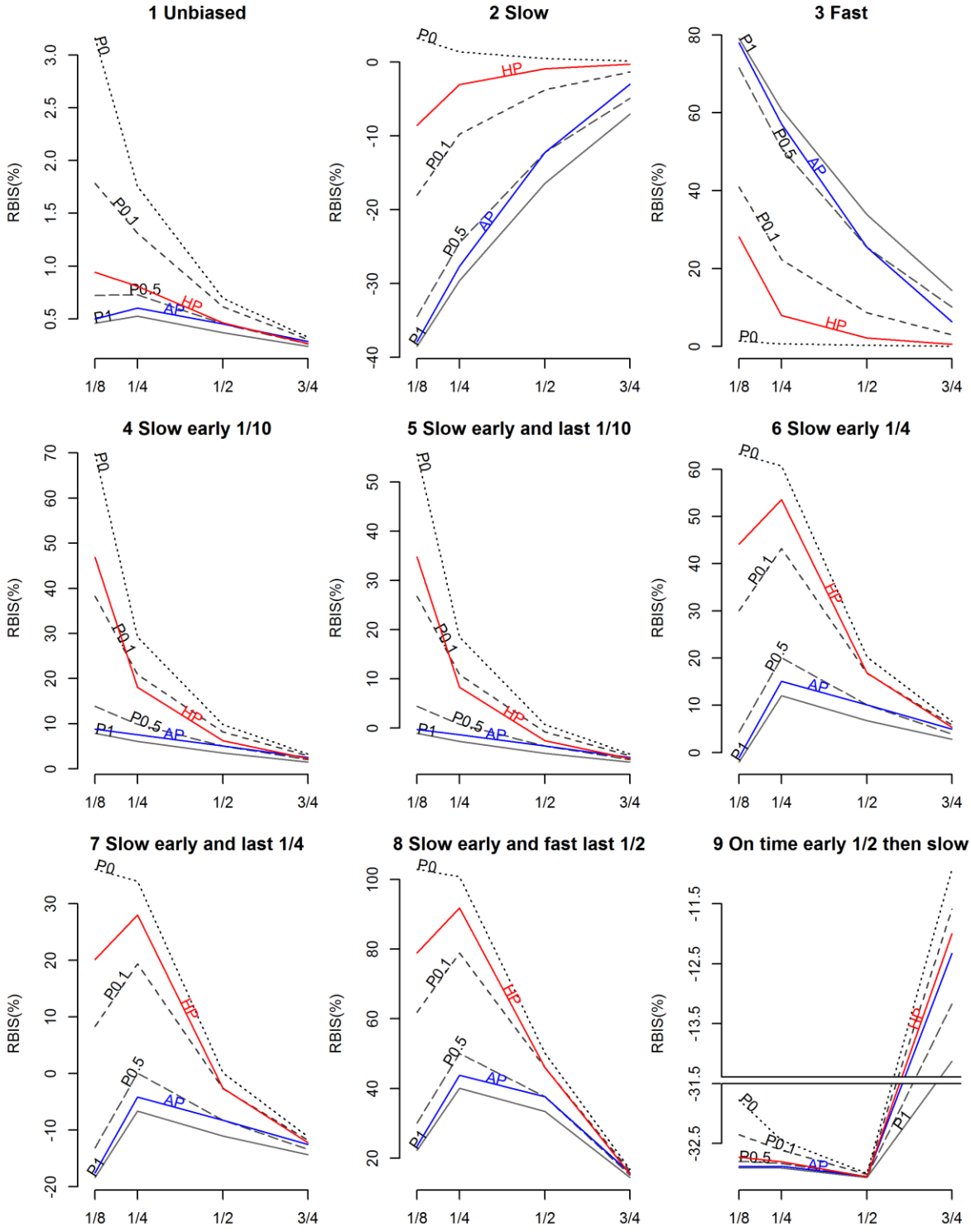


Figure 3-6. The RBIAS of each simulation study using the various methods (P0, P0.1, P0.5, AP, HP) when recruited the first 1/8, 1/4, 1/2 or 3/4 of the subjects. As the results for study 9 *On time early 1/2 then slow* are too close, their labels are overlapped.



CHAPTER FOUR:

Open Source R Code and Smart Phone Application for Bayesian Accrual Prediction for Interim Review of Studies

(To be submit to *Clinical Trials)*

(The references are formatted for APA style)

Abstract

Background: Subject recruitment for medical research is challenging. Slow patient accrual leads delays in research. Researchers need reliable tools to manage the accrual rate.

Methods: Our previously developed Bayesian method integrates researchers' experience on previous trials and data from the current study, providing reliable prediction on accrual rate for clinical studies.

Results: In this paper, we present a friendly graphical user interface program developed in R which can be easily used by clinical researchers. A closed form solution for the total subjects that can be recruited in a fixed time is also derived and built in an Android system using Java, which can be used by a web browser and Smartphone carriers.

Conclusions: This application provides a more convenient platform for estimation and prediction of the accrual process.

Key words: Subject Accrual, Bayesian Methods, Smartphone Application, Statistical Software

1. Introduction

Subject recruitment is critical and challenging in medical research. Researchers tend to overestimate the pool of available subjects and underestimate the time they need to obtain the proposed sample sizes for their research studies. This is known as the "Lasagna's Law" "(Lasagna et al., 1979) and "Muench's Third Law" (Bearman et al., 1974). Studies have shown that more than 80% of clinical trial studies ran longer than their accrual goals (van der Wouden, et al., 2007). The delayed subjects recruitment and/or insufficient sample size would have serious deleterious consequences. Extending the recruitment time frame will lead to increased costs and usage of resources. If the proposed sample size is not achieved, the study may be seriously underpowered. Therefore, it is important for researchers to monitor the accrual all along their study.

A number of studies have been done to model and predict patient accrual process. Both Barnard (2010) and Zhang (2012) recently reviewed the prediction methods. Barnard (2010) summarized the current accrual models into five categories: (a) unconditional model, (b) conditional model, (c) Poisson model, (d) Bayesian model, and (e) Monte carlo simulation Markov model. Zhang (2010) further compared the Poisson process-based models. Among the proposed methods, both Barnard and Zhang addressed the Bayesian methods conducted by Gajewski (2008). In addition, the Bayesian approach can utilize researcher's previous experience in similar studies or clinical opinion and incorporate them into prior knowledge. When actual accrual data are available, the predictive distribution of the accrual becomes the weighted average of the prior distribution and the actual observed data. As more data are collected, the weight on currently observed data will be increased while the weight of prior information will be

decreased. The method provide an objective assessment of the accrual process. If the predicted accrual is so slow that it threatens the promised sample size and increases the trial duration of a study, this monitoring will allow internal review boards (Schroen, 2010) to suggest mid-course corrections in the trial, such as adding additional centers to a multi-center study, hiring additional study coordinators to broaden the search for volunteers, updating the inclusion/exclusion criteria etc. On the other hand, the method can also prevent a researcher from overreacting to slow accrual in the beginning of the study. If the accrual is faster than planned, the prediction model can provide an estimated closure date to avoid unnecessary patient recruitment.

As most clinical researchers are not proficient in using the algorithm in the Bayesian method to do the accrual prediction, it will be convenient for them to have simple and easy to use interface. In this paper, we present our R *accrual* package, which only requires the researchers to input the original design information and the updated accrual data with simple click and point (R Development Core Team, 2012), while using the Bayesian prediction model calculated in the background (Gajewski, Simon and Carson, 2008). The design information includes total time proposed and time subjects proposed, which are usually required for Institutional Review Board (IRB). The proposed software has three major functions: (a) provide the estimate of the total number subjects that the trial will have recruited within the planned time frame, (b) provide the time frame that the trial will successfully recruit enough number of subjects, and (c) produce diagnostic panel plots of the actual accrual data, such as the cumulative accrual plot, the distribution of the accrual etc. The *accrual* package has been promoted on the R listserv,

and is ready to be used by both statisticians and clinical researchers in evaluating and monitoring their subject accrual.

In our recent study (Jiang, Simon, Mayo & Gajewski, accepted), we derived a closed form solution for the posterior prediction of the accrual. The duration for the remaining number of subjects needed to be accrued is distributed as inverse beta. The percentile of the duration then can be calculated using the normal approximation of beta distribution. We also derive a closed form solution for the posterior prediction of the remaining subjects that can be recruited in a fixed time, which is a negative binomial. As shown in the appendix, the negative binomial can be approximated using a normal distribution. Based on the closed form algorithm, we developed a web browser and an android version of the accrual calculator, which can be easily installed and used by a smartphone carrier.

2. Example using the R *accrual* package

The R *accrual* package includes an example data set, three major functions described below, and a graphical user interface that provides menu driven access to these functions in R (Figure 1). The major three R functions are *accrual.n*, *accrual.T* and *accrual.plots*. The function *accrual.n* calculates the prediction of the number of patients to be recruited in fixed time. The function *accrual.T* predicts the time to reached targeted sample size. The function *accrual.plots* provides a panel of plots for data diagnostics. The *accrual.gui* provides an interface that the users can choose any of the three options as needed. We use an example of clinical trial to illustrate how to use the R package and interpret the results.

Suppose in a clinical trial, the researcher's original proposal is to recruit 300 patients in 3 years (36 months). Assuming that the investigator is 50% confident that the accrual can be done within the planned time, that is P=0.5, and that the trial has not been started, the estimation of total patients will be recruited in three years can be done using the function *accrual.n*, *accrual.n (n=300,T=36,P=0.5,m=0,tm=0).*In addition, users can also choose to use the GUI window. Figure 2A shows the interface window for the users to estimate "How many patients will you recruit?", which functions exactly as *accrual.n*. In this case, the "Total sample size" is 300. "Targeted finish time in months" is 36. The researcher can choose any confidence level (0 to 1) by using a slider or directly entering the confidence level in the blank directly. In the current case, the trial has not been started, therefore both "Subject recruited" and "Total months after started" are 0. Through either of the two approaches, the prediction of the accrual is shown in Figure 2B. The white line is the estimate of the prediction, with the grey tunnel as the prediction intervals. The histogram of estimated total accrual in 36 months is shown on the right. The horizontal line indicates the target sample size. On the top left corner of the figure, there are input information and the summary of the results. In this example, there will be 300 subjects recruited in 36 months, with 95% prediction interval (247, 361). It will take 35.9, with 95% prediction interval (29.9, 43.8) months for the investigators to recruit 300 subjects. The plots and the summary results will help the investigators and the IRBs to monitor and predict the progress of the clinical trial. Please be aware there maybe a slight difference in the results between the displayed results and what will be obtained by the reader using the package. This is due to the using of simulation approach in the

calculation. The trivial difference will not affect the interpretation of the results and the evaluation of the trial progress.

Using the R accrual package, the recruitment of the trial can be monitored all along the process. For example, if the trial is in progress with 75 subjects recruited in 12.98 months, the prediction can be done as either use *accrual.n* function or the interactive R window for data input, as shown in Figure 3A. Figure 3B shows the corresponding accrual plot and summary of the results. If 1/4 of total sample size, that is 75 subjects, recruited in 12.96 months, then the predicted total subjects can be recruited in 3 years is 241 with 95 prediction interval (210, 277). The results indicate that it is highly probable that the study will not be able to recruit enough subjects within the study time frame given the current recruitment process. The investigator should consider strategies to increase the accrual rate, such as adding one more study center, and changing the study protocols. The researchers can use this function/option throughout their trial. The results also help the IRBs to evaluate the progress of the study objectively.

In addition to estimating the total number of subjects in a fixed time, the investigators may also be interested in estimating the time frame to finish recruiting a certain number of subjects. It can be done by using function *accrual.T*, or the second option (Figure 1) "How long will it take to reach the targeted sample size?". Similarly as the previous example, the researcher's original proposal is to recruit 300 patients in 36 months, and the researcher's confidence level is also 0.5. The R window interface and output are shown in Figure 4A and Figure 4B respectively. In Figure 4B, the summary of input information and output are shown on the left bottom corner. The vertical line shows the targeted time (36 months). As shown clearly in Figure 4B, the predicted time to finish

recruiting 300 subjects is 44.1 months, with 95% predicted interval (38.9, 50.3). The predicted accrual is much slower than planned 36 months. The results indicated that both the IRBs and the investigator should pay special attention to the accrual rate as discussed above.

The Bayesian accrual model is based on the assumption that the distribution of waiting time is exponential and the rate of the accrual is constant. Violations of the assumption may lead to biased estimation. Therefore, it is useful to check whether data (w) meet the assumption and is suitable for the current method via option "Diagnostic Panel". Figure 5 A shows the windows interface of this option, in which the researchers can load the raw time gap data through point and click. Figure 5B shows four figures that help to understand the data distribution. The figure on the top left is the exponential quantile plot, which checks whether the distribution of waiting times is exponential. The current plot shows that data are off from the straight line. The top right figure shows the histogram of the waiting times, where the red line is the theoretical exponential distribution. The figure of waiting time verse cumulative accrual time is shown on the bottom left, and the figure of total accrual verse cumulative accrual time is shown on the bottom right. Both of the graphs show that this trial is piecewise constant with slower accrual in the beginning and at the end.

3. Using the web-based calculator and smartphone applications

As discussed in the paper (Gajewski, Simon and Carson, 2012), the closed form of time frame of accrual shows to be distributed as inverse beta. As we know, we can use normal approximation for both of the beta distribution, which accelerate the speed of

calculation greatly. The normal approximation algorithm is adopted in Java and used for the estimations in the accrual. Our group developed a web-based accrual software using Java. An example of using this web calculator is shown in Figure 6. The link to the software can be found at http://biostat-pts.kumc.edu/velos/RPackages/Home.html.

With the development and wide use of smartphone devices, the accrual smartphone application, compared to R packages or web-based application, is more convenient and easier to clinical researchers. Using closed form solution for Bayesian accrual model and normal approximation, the methods are adopted into an Android application using Java. Figure 8 shows the use of an Android phone in the process of accrual monitoring.

4. Feedback from clinical investigators

In order to evaluate the developed software, we demonstrated the software to clinical investigators at our academic medical center. They were given a single item: "I would recommend this software to other researchers." There are seven response options for this item: 1) `Strongly disagree'; 2) `Disagree'; 3) `Somewhat disagree'; 4) `agree or disagree'; 5) `Somewhat agree'; 6) `Agree'; 7) `Strongly agree'. We received eight responders out of 16 investigators, with three researchers chose `Strongly agree', two researchers choosing `Agree', one choosing `Somewhat agree' and only one choosing `Somewhat disagree'. The mean score of the answer is 5.87, which indicates agreement that the software is useful and should be introduced to more investigators.

5. Discussion

Both the R *accrual* package, the web-based as well as the Smartphone application for patient accrual, are based on the assumption that the accrual is constant. However, as

we know, in the real situation, the accrual is not constant or only piecewise constant, such as slow in the beginning and/or in the end. Although the software is robust when the assumption is only slightly violated (Jiang, Simon, Mayo & Gajewski, accepted), is still better to check the data distribution and use it with caution when the assumptions are obviously violated. Future work is to continue assessing the assumptions, build models when needed, and translate these methods to easy to use software for monitoring of clinical trials.

Acknowledgments

This work was supported by a NIH grant from NCI awarded to The University of Kansas Cancer Center (The KUCC) #1 P30 CA168524.

Figure 4-1. The main menu of R accrual Package with three options.

Figure 4-2. An example of using R accrual package to calculate the number of patients can be recruited in the beginning of clinical trial. (A) The interactive R window for data input (B) The R output for summarized results and accrual plot.

(A)

(B)

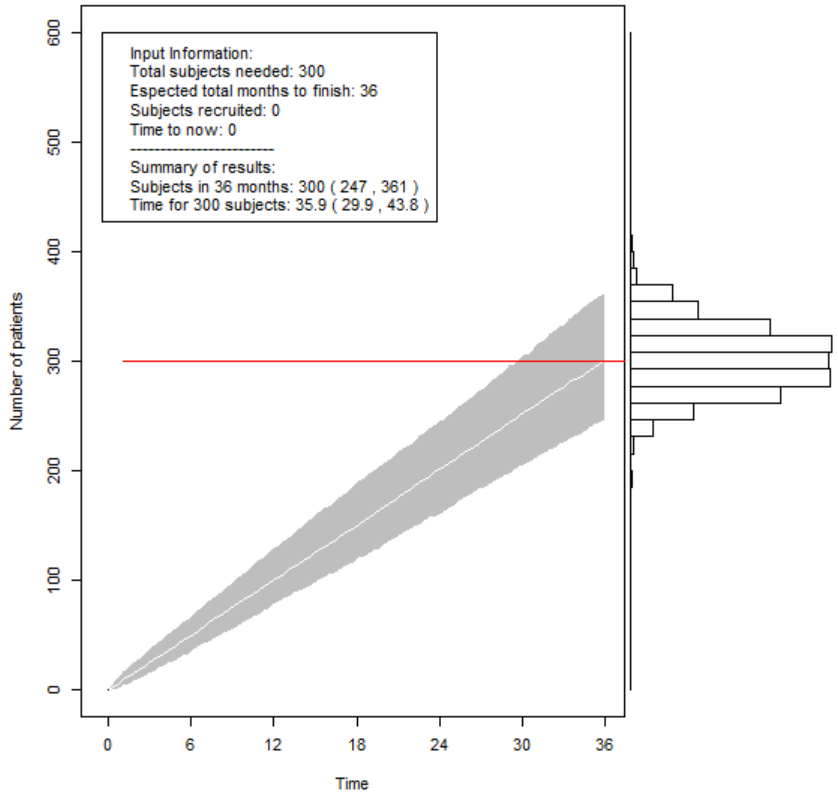
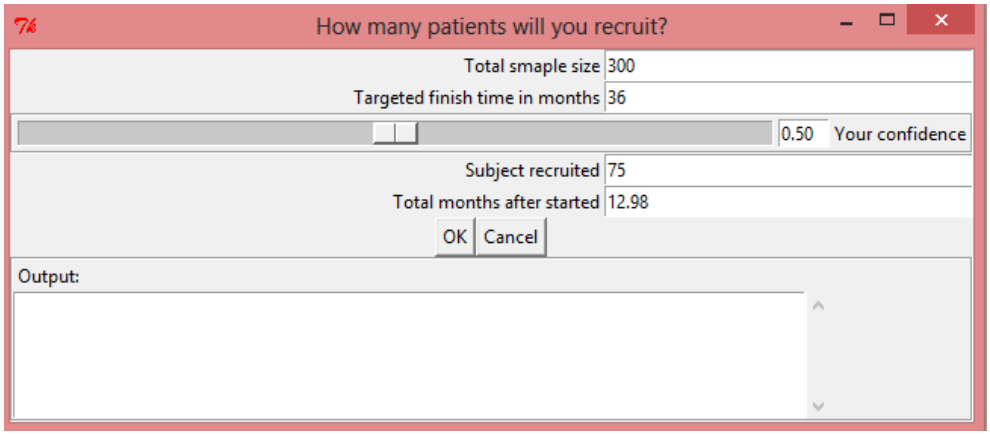


Figure 4-3. An example of using R *accrual* package to calculate the number of patients can be recruited when 75 subjects has been recruited. (A) The interactive R window for data input (B) The R output for summarized results and accrual plot.

(A)



(B)

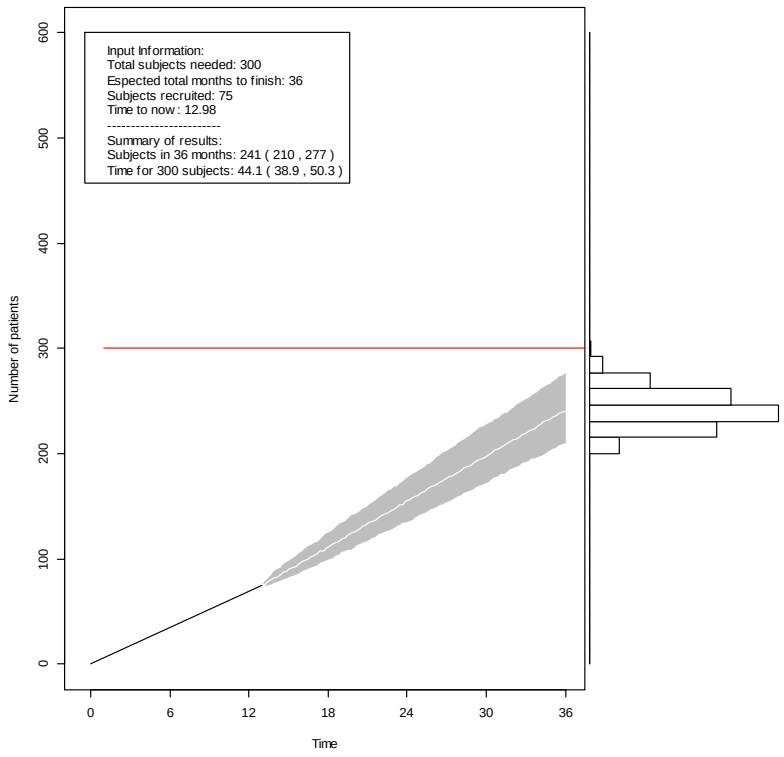
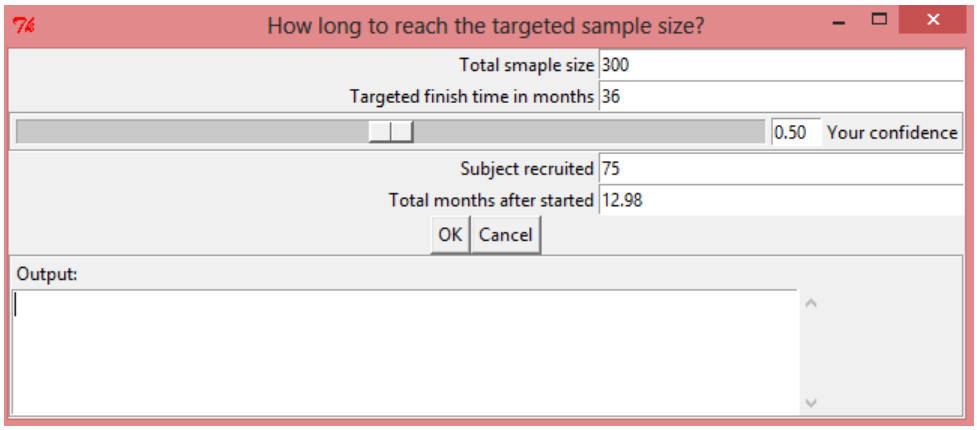


Figure 4-4. An example of using R *accrual* package to calculate the time frame to reach the targeted sample size when 75 subjects has been recruited. (A) The interactive R window for data input (B) The R output for summarized results and accrual plot.

(A)



(B)

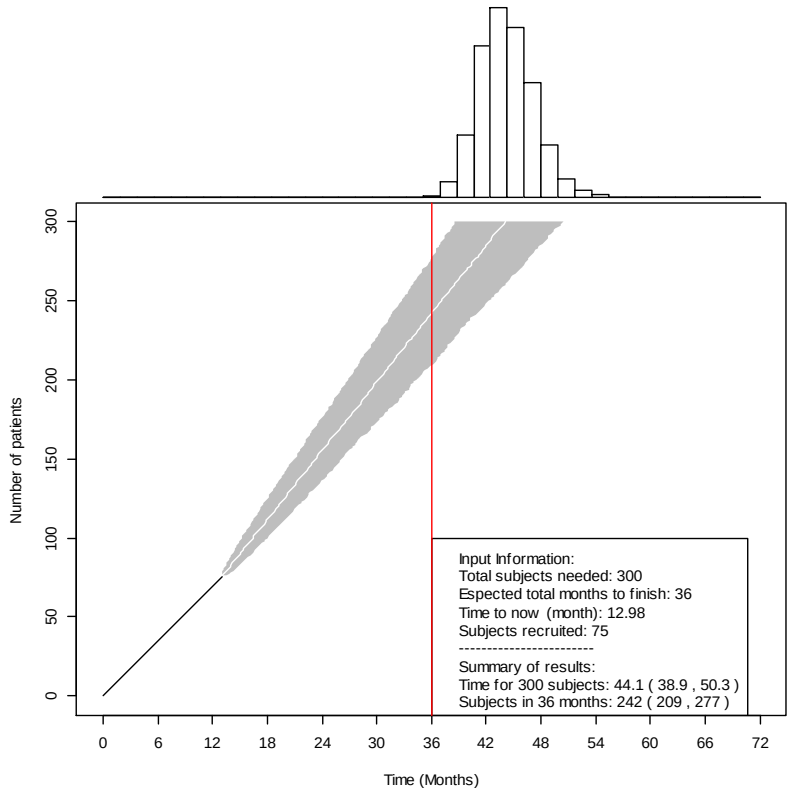
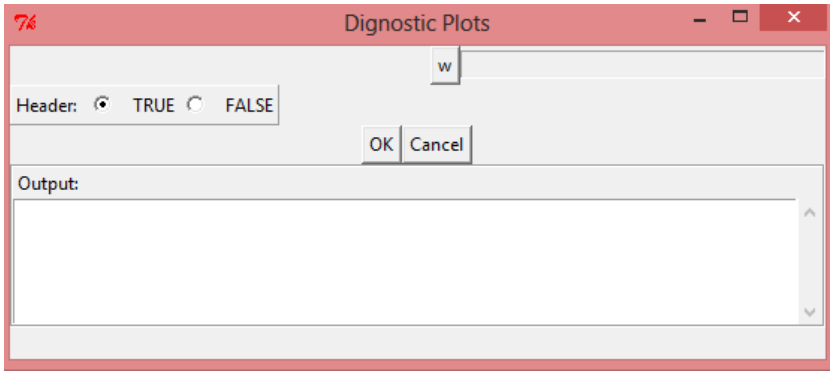


Figure 4-5. An example of using R *accrual* package to check the distribution of accrual data (A) The interactive R window to input data (B) The R output for exponential quartile plot for waiting times (top left), the histogram of the individual waiting times (top right), waiting times verse cumulative accrual time (bottom left), and the number of subjects verse cumulative accrual time (bottom right)

(A)



(B)

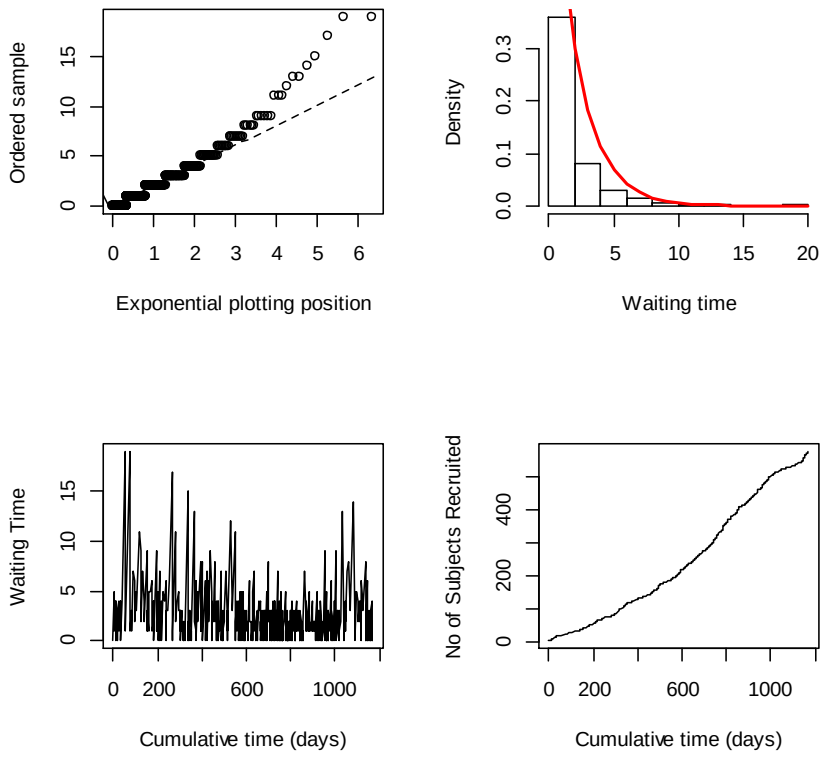


Figure 4-6. An example of using accrual web based software to calculate the number of patients can be recruited when 1/4 of the projected subjects has been recruited.

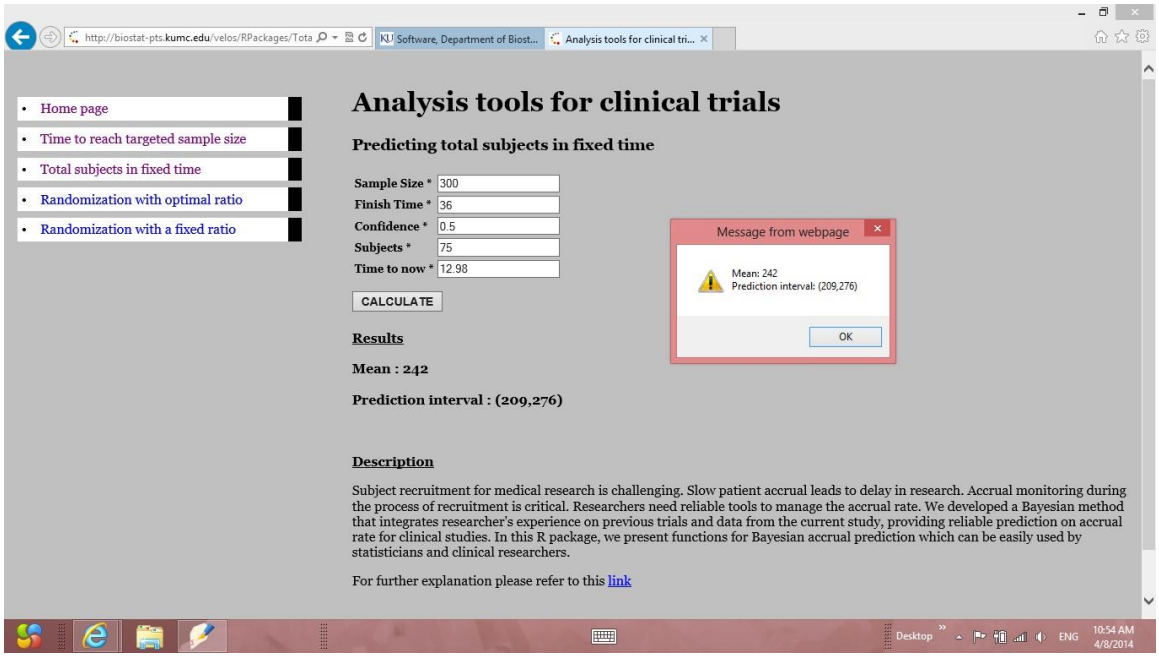


Figure 4-7. An example of using accrual web based software to calculate the time frame to reach the targeted sample size when 1/4 of the projected subjects has been recruited.

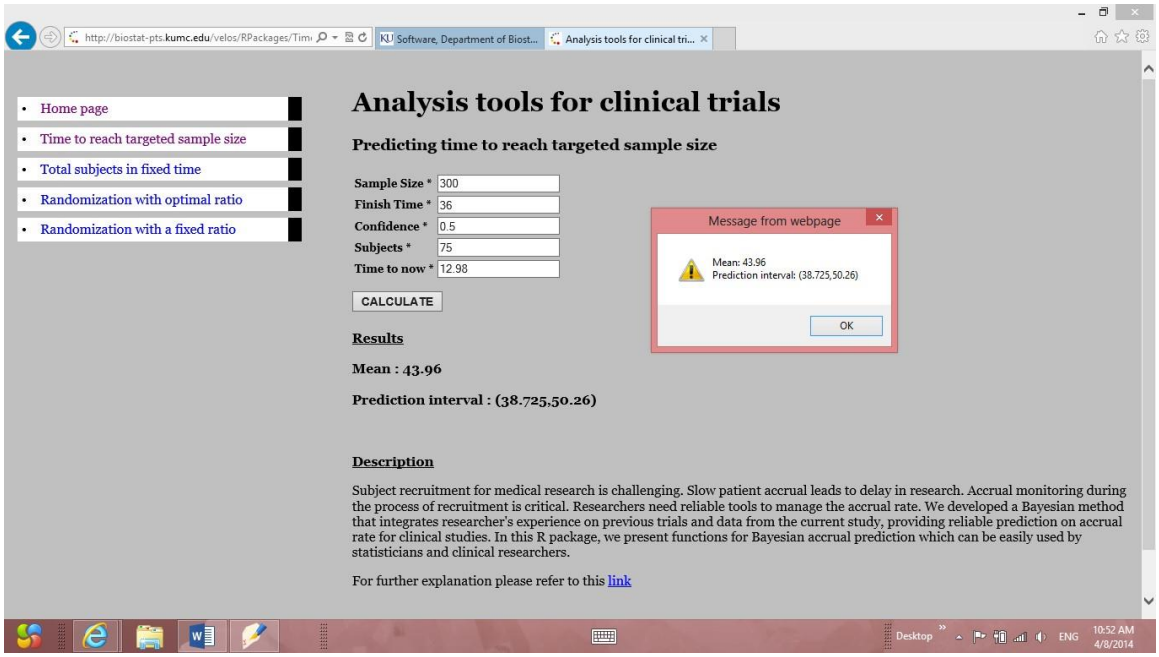
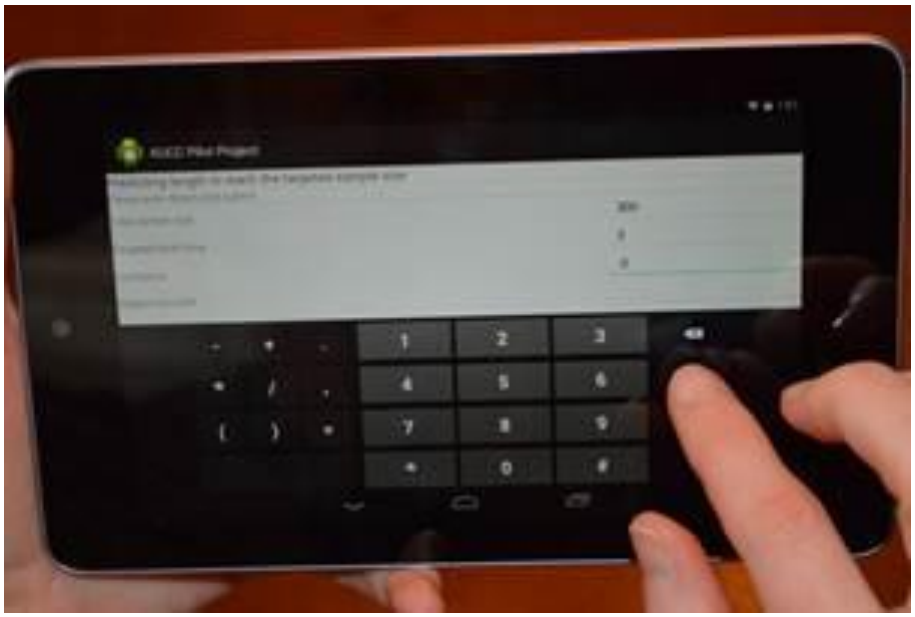


Figure 4-8. The using of accrual Smartphone application.



Appendix

Using R, the Bayesian method for simple accrual (Gajewski, 2008) can be easily done using simulations. To speed up the calculation, it is better to develop closed form that can be used in Java. Assuming the investigator planned to recruit *n* subjects in *T* days in the original protocol. The Bayesian model assumes that the waiting time (*w*) for each successive patient follows an exponential distribution, that is $w_i \sim \exp(\theta)$, where θ represents the average accrual time for the *i*th subject. We also assume that distribution of θ is inverse gamma, $\theta \sim IG(nP, TP)$, where *P* is the investigator's confidence on the original plan on a 0-1 scale. In the process of a trial, suppose m subjects have been collected in T_m ($T_m = \sum_{i=1}^{m} w_i$) time period. The posterior distribution for θ is

$$f(\theta|m, T_m) = \frac{(TP+T_m)^{nP+m}}{\Gamma(nP+m)} \theta^{-(nP+m+1)} e^{-\frac{TP+T_m}{\theta}} \quad (1)$$

For fixed T, assuming the rest of subjects can be recruited are $\eta \sim Poi(\frac{T-Tm}{\theta})$, then the posterior predictive distribution of η is

$$g(\eta) = \int_0^\infty \frac{(TP+T_m)^{nP+m}}{\Gamma(nP+m)} \theta^{-(nP+m+1)} e^{-\frac{TP+T_m}{\theta}} \frac{(\frac{T-Tm}{\theta})^\eta}{\eta!} e^{\frac{T-Tm}{\theta}} d\theta \quad (2)$$

$$= \frac{(TP+T_m)^{nP+m}(T-Tm)^\eta}{\Gamma(nP+m)\eta!} \int_0^\infty \theta^{-(nP+m+\eta+1)} e^{-\frac{TP+T}{\theta}} d\theta$$

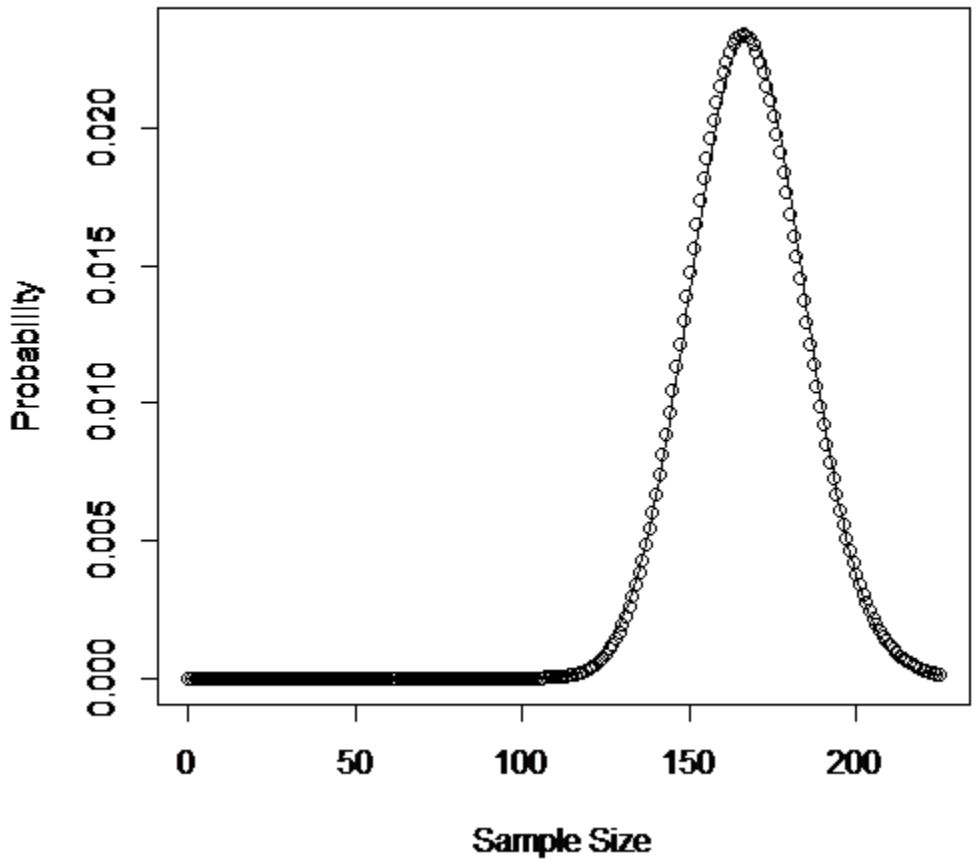
$$= \frac{(TP+T_m)^{nP+m}(T-Tm)^\eta \Gamma(\mathrm{nP+m+\eta})}{(TP+T)^{\mathrm{nP+m+\eta}} \Gamma(nP+m)\eta!} \quad (3)$$

Define, $p = \frac{TP+Tm}{\mathrm{TP+T}}$, and $r = nP + m$, then $g(\eta) = p^r (1-p)^\eta \frac{\Gamma(\mathrm{r+\eta})}{\Gamma(r)\eta!}$. (4)

The above formula shows that the distribution of η is negative binomial, NG (r,p). When the number of successes *r* is large, and *p* is neither very small nor very large, NG (r,p). can be approximated with $N(\frac{r(1-p)}{p}, \frac{r(1-p)}{p^2})$. For example, when n=300, T=36,

P=0.05, m=75. Tm=12.98, p is 0.5737037. Obviously, it is not too small or too large, and r is 225, which is large. Figure S1 shows the density NG (225, 0.5737037) in dots, with the solid line is N(167.1885, 291.4196). The graph shows that it is reasonable to use normal approximation in estimating of η, the rest of subjects can be recruited. Then the total number of subjects can be recruited in fixed T time is η +T, The calculation of normal quantile can be easily adopted in Java.

Figure 4S-1. The distribution NB (225, 0.5737037) (dotted line) and its corresponding Normal distribution approximations (solid line)



CHAPTER FIVE:

SUMMARY AND FUTURE DIRECTIONS

Use of indirect measurement information in the form of a prior distribution is becoming more popular (Berry, 2004) and makes Bayesian methods more efficient (Samaniego & Reneau, 1994). We applied Bayesian methods to estimate construct validity in patient reported outcomes and in monitoring of subjects' accrual in clinical trials.

In the study of patient reported outcomes, we developed and evaluated Bayesian Instrument Development (BID). In BID, the prior probablility distribution is derived from the knowledge of content validity. Participant data is then used to update the posterior distribution. To test the stability of BID, we used simulations to compare BID with classical instrument development procedures. The comparisons were conducted under three different conditions: the priors (experts) are all unbiased, the priors (experts) are all biased, and the priors are contaminated. Our results showed that BID performs better than traditional factor analysis when using a flat or "good informative" prior. The results are consistent with findings by Lee (1981) and Lee & Shi (2000). It also confirms Samaniego and Reneau's (1994) general findings of the Bayesian and frequentist comparison: a Bayesian estimator is always superior to a frequentist estimator if the bias in the Bayesian prior is smaller than the sample standard deviation.

In the study of patient accrual for clinical trials, we developed two hierachical extensions to the Bayesian constant accrual model (Gajewski, Simon & Carlson, 2008). In brief, the researchers' previous experience is incorporated into the prior distribution, and then the accrual data are used to update to a posterior distribution for the estimation of accrual duration. The two new extensions we proposed are an accelerated prior and a hedging prior. The performance of the Bayesian constant accrual model and the extension

were evaluated by clinical trial data and through simulation. Overall, the results showed that informative priors perform well when accrual is on target or slightly off (the prior is unbiased or slightly biased), but are worse when accrual is off target (the prior is biased). The flat or weakly informative prior performs well when the prior is biased, but are less efficient when the prior is unbiased. As discussed by Samaniego and Reneau (1994), bias in the Bayesian prior should be controlled under a certain level to be more efficient. Our proposed hedging prior adds the similarity between the researchers' experience and the current data as a "control" in the utilization of prior information. It weighs the prior information more heavily in the estimation of the posterior mean when the prior is unbiased, but less otherwise. In our study, the hedging prior performs much like the weak priors when we have a biased prior, but closer to the strong informative priors when we have an unbiased prior.

For both methods discussed above, we developed software with a graphical user interface that can be easily operated by non-statisticians, most of whom are not familiar with Bayesian statistics. BID is based on R and WINBUGS, and the BID software has been demonstrated to nursing researchers. For patient accrual, the R package *accrual* has been submitted and accepted by R CRAN, and the package is ready to used. We demonstrated the software to clinical researchers in a medical center. The demonstration showed that the researchers agree that the software is useful and should be introduced to more investigators.

Overall, our Bayesian statistical methods development helps decrease the cost and resource utilization of health care studies. Using BID, the researcher can eliminate the need for unnecessary continuation of data collection for larger samples as required by the

classical instrument development approach, which will in turn decrease the time and cost of the study. Monitoring accrual in clinical trials will allow the investigator and the interal review boards to evaluate trial progress objectively and suggest mid-course corrections for the trial if accrual is slower than expected, avoiding delays in trial progress and increases in cost.

A few suggestions for future studies are listed as follows:

(1) In BID, participants' data are assumed to continuous..Many clinical questionnaires are in ordinal or binary form. It is necessary to develop Ordinal Bayesian Instrument Development (OBID) with an item response theory model (Albert & Chib, 1993; Beck & Gable, 2001).

(2) In the study of patient reported outcomes, we assume the experts' opinion on items is independent. However, the experts' opinions on items are most likely correlated in real situations. Therefore, a correlated model may be more suitable for estimation of the content validity (Albert et al., 2012). Each expert will have an individualized prior based on their prior belief. This hierarchical approach will make it possible for us to evaluate the effect of a single expert on the analysis and draw conclusions (Ansari, Jedidi, & Jagpal, 2000).

(3) In the Bayesian constant accrual model, the accelerated prior performs just like an informative prior when only the first half of the data is collected. A stronger degree of acceleration, such as a cubed acceleration, $P = \left(1 - \frac{m}{n}\right)^3$, may better compromise between prior information and data.

(4) Many clinical trials are conducted in different places (i.e., multi-centered). The current model does not include variation from different centers. In the future, we

can use a hierarchical model with each center having a specific prior character which may represent the potential subject recruitment ability of that center (Zhang & Long, 2012).

(5) As previously discussed, the actual accrual processes are typically not constant, but piecewise constant. For example, accrual is usually slow for the first few subjects as there is a training period for the staff. In this case, we can use a finite mixture model hyper prior to better estimate the piecewise constant case (McLachlan & Peel, 2004).

References:

1. Abbas, I., Rovira, J., & Casanovas, J. (2007). Clinical trial optimization: Monte Carlo simulation Markov model for planning clinical trials recruitment. *Contemporary clinical trials*, 28(3), 220-231.
2. Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., & Rousseau, J. (2012). Combining expert opinions in prior elicitation. Bayesian Analysis, 7(3), 503-532.
3. Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. Journal of the American statistical Association, 88(422), 669-679.
4. Albert, J. (2007). *Bayesian computation with R*. New York: Springer.
5. Alonso,A., Laenen, A., Molenberghs, G., Geys, H., & Vangeneugden, T.(2010). A Unified Approach to Multi-item Reliability. *Biometrics*, *66*, 1061-1068.
6. Anisimov, V. V., & Fedorov, V. V. (2007). Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in medicine*, 26(27), 4958-4975.
7. Ansari, A., Jedidi, K., & Jagpal, S. (2000). A hierarchical Bayesian methodology for treating heterogeneity in structural equation models. Marketing Science, 19(4), 328-347.
8. Bakhshi, A., Senn, S., & Phillips, A. (2013). Some issues in predicting patient recruitment in multi‐centre clinical trials. *Statistics in medicine*, 32(30), 5458-5468.
9. Barnard, K. D., Dent, L., & Cook, A. (2010). A systematic review of models to predict recruitment to multicentre clinical trials. *BMC medical research methodology*, 10(1), 63.

10. Bates, D., & Sarkar, D. (2012) R port: http://netlib.bell-labs.com/netlib/port/ accessed on April.
11. Beck, C. T., & Gable, R. K. (2001). Item response theory in affective instrument development: An illustration. *Journal of nursing measurement*, 9(1), 5-22.
12. Bearman, J. E., Loewenson, R. B., & Gullen, W. H. (1974). Muench's postulates, laws and corollaries, or biometrician's views on clinical studies (Biometric Note 4). Bethesda (MD): Office of Biometry and Epidemiology, National Eye Institute, National Institutes of Health.
13. Berry, D.A.(2006), Bayesian clinical trials. *Nat Rev Drug Discov*. 5(1):27-36.
14. Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3), 385-402.
15. Breau, R. H., Carnat, T. A., & Gaboury, I. (2006). Inadequate statistical power of negative clinical trials in urological literature. *The Journal of urology*, 176(1), 263-266.
16. Broemeling, L.D. (2007). *Bayesian Biostatistics and Diagnostic Medicine*. Boca Raton, FL: Chapman & Hall/CRC.
17. Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis 1*: 473-550.
18. Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Pacific Grove, CA: Duxbury.
19. Carter, R. E. (2004). Application of stochastic processes to participant recruitment in clinical trials. *Controlled clinical trials*, 25(5), 429-436.

20. Carter, R. E., Sonne, S. C., & Brady, K. T. (2005). Practical considerations for estimating clinical trial accrual periods: application to a multi-center effectiveness study. *BMC medical research methodology*, 5(1), 11.

21. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., …Hays, R. D. on behalf of the PROMIS Cooperative Group. (2010). Initial item banks and first wave testing of the Patient–Reported Outcomes Measurement Information System (PROMIS) network: 2005–2008. *Journal of Clinical Epidemiology*, *63*, 1179-94.

22. Chaloner, K. (1987). A Bayesian Approach to the Estimation of. Variance Components for the Unbalanced. One-Way Random Model, *Techn. 29*:322-337.

23. Cohen, J. (1988), *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

24. Cook, J. D., Jairo, A., & Pericchi, L. R. (2011). Skeptical and Optimistic Robust Priors for Clinical Trials. UT MD Anderson Cancer Center Department of Biostatistics Working Paper Series. Working Paper 65.

25. Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Boston, MA: Harcourt, Brace, Jovanovich.

26. Cupertino, A. P., Hunt, J. J., Gajewski, B. J., Jiang, Y., Marquis, J., Friedmann, P. D., ... & Richter, K. P. (2013). The index of tobacco treatment quality: development of a tool to assess evidence-based treatment in a national sample of drug treatment facilities. *Substance abuse treatment, prevention, and policy*, 8(1), 13.

27. Duan, Y., Ye, K., & Smith, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17(1), 95-106.

28. Deshpande, P. R., Rajan, S., Sudeepthi, B. L., & Nazir, C. P. (2011). Patient-reported outcomes: A new era in clinical research. Perspectives in clinical research, 2(4), 137-44.
29. Efron,B, (2010), The future of indirect evidence, *Statistical Science*, *25*: 145–157.
30. Eubank, R. L., & Kupresanin, A. (2011). *Statistical Computing in C++ and R*. CRC Press.
31. Fúquene, J. A., Cook, J. D., & Pericchi, L. R. (2009). A case for robust Bayesian priors with applications to clinical trials. Bayesian Analysis, 4(4), 817-846.
32. Gajewski, B.J., Sedwick, J.D., & Antonelli, P.J. (2004). A log-normal distribution model of the effect of bacteria and ear fenestration on hearing loss: A Bayesian approach. *Statistics in Medicine*, *23*, 493-508.
33. Gajewski, B.J., Hart, S., Bergquist, S., & Dunton, N. (2007). Inter-rater reliability of pressure ulcer staging: Ordinal probit Bayesian hierarchical model that allows for uncertain rater response. *Statistics in Medicine*, *26*, 4602-4618.
34. Gajewski, B. J., Simon, S. D., & Carlson, S. E. (2008). Predicting accrual in clinical trials with Bayesian posterior predictive distributions. Statistics in Medicine, 27(13), 2328-2340.
35. Gajewski, B.J., Nicholson, N., & Widen, J.E. (2009). Predicting hearing threshold in non-responsive subjects using a log-normal Bayesian linear model in the presence of left censored covariates. *Statistics in Biopharmaceutical Research*, *1*, 137-148.
36. Gajewski, B.J., Coffland, V., Boyle, D., Bott, M.J., Price, L., Leopold, J., & Dunton, N. (2012). Assessing content validity through correlation and relevance tools: A

Bayesian randomized equivalence experiment, *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 8*, 81-96.

37. Gajewski, B. J., Simon, S. D., & Carlson, S. E. (2012). On the existence of constant accrual rates in clinical trials and direction for future research. *International journal of statistics and probability*, 1(2), p43.
38. Gajewski, B.J., Price, L., Coffland, V., Boyle, D., & Bott, M.J. (2013). Integrated analysis of content and construct validity of psychometric instruments. *Quality & Quantity*, 47(1), 57-78.
39. Garthwaite, P.H., Kadane, J.B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association, 100*, 680-701.
40. Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian data analysis*. New York: Chapman and Hall.
41. Ghosh, M. (2011). Objective priors: An introduction for frequentists. Statistical Science, 26(2), 187-202.
42. Gilks, W.R., Thomas, A., & Spiegelhalter, D.J. (1994). A language and program for complex Bayesian modeling, *Statistician, 43*, 169 –177.
43. Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. Bayesian Analysis, 1(3), 403-420.
44. Green, S.B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika,74*, 121-135.
45. Grosjean, P. S. R. (2013). A GUI API for R. UMONS, Mons, Belgium.

46. Grosjean, P. (2012). SciViews: A GUI API for R. UMONS, Mons, Belgium. URL http://www.sciviews.org/SciViews-R.

47. Haidich, A. B., & Ioannidis, J. P. (2001). Determinants of patient recruitment in a multicenter clinical trials group: trends, seasonality and the effect of large studies. *BMC medical research methodology*, 1(1), 4.

48. Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., & Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3), 1047-1056.

49. Hoffmann, T. J., & Laird, N. M. (2009). fgui: A Method for Automatically Creating Graphical User Interfaces for Command-Line R Packages. *Journal of Statistical Software*, 30(2), 1-14.

50. Hughes, M. D. (1991). Practical reporting of Bayesian analyses of clinical trials. *Drug information journal*, 25(3), 381-393.

51. Ibrahim, J. G., & Chen, M. H. (2000). Power prior distributions for regression models. *Statistical Science*, 46-60.

52. Jaykaran, D. S., Yadav, P., & Kantharia, N. K. (2011). Negative studies published in Indian Medical Journals are underpowered. *Indian Pediatr*, 48, 490-1.

53. Jiang, Y., Simon, S., Mayo, M.S. & Gajewski, B.J. (Accepted) Performance of Constant Accrual Model and Alternatives on Clinical Data and Simulation. Statistics in Medicine.

54. Johnson, N. L., Kotz, S.,& Balakrishnan N (1995). Chapter 25 of Continuous Univariate Distributions Volume 2 (2nd Edition) Wiley: New York.

55. Kadane, J., & Wolfson, L. J. (1998). Experiences in elicitation. Journal of the Royal Statistical Society: Series D (The Statistician), 47(1), 3-19.

56. Kadane, J. B. (2006). Is ‘‘objective Bayesian analysis’’objective, Bayesian, or wise?(comment on articles by Berger and by Goldstein). Bayesian Anal, 1(3), 433-436.

57. Keen, H. I., Pile, K., & Hill, C. L. (2005). The prevalence of underpowered randomized clinical trials in rheumatology. *The Journal of rheumatology*, 32(11), 2083-2088.

58. Laplace P (1774). "Memoire sur la Probabilite des Causes par les Evenements." l'Academie Royale des Sciences, 6, 621-656.

59. Lasagna, L. (1979). Problems in publication of clinical trial methodology. *Clinical pharmacology and therapeutics*, 25(5 Pt 2), 751-753.

60. Lee, S.Y. (1981) A Bayesian approach to confirmatory factor analysis. *Psychometrika*, *46*, 153-160.

61. Lee, S.Y., & Shi, J.Q. (2000). Joint Bayesian analysis of factor scores and structural parameters in the factor analysis model. *Annal of the Institute of Statistical Mathematics*, 52, 722-736.

62. Lee, S.Y., & Song, X.Y. (2004) Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, *39*, 2004, 653-686.

63. Lee, S.Y. (2007). *Structural Equation Modeling: A Bayesian Approach*. Wiley, New York.

64. Lunn, D.J., Thomas, A., Best, N., & Spiegelhalter, D. (2000) WinBUGS--a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*: 325-337.

65. Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25), 3049-3067.

66. McGrayne, S. B. (2011). The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy. Yale University Press.

67. McLachlan, G., & Peel, D. (2004). Finite mixture models. John Wiley & Sons.

68. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillian.

69. Michalak, E.E., & Murray, G. (2010). Development of the QoL.BD: a disorder-specific scale to assess quality of life in bipolar disorder. *Bipolar Disorder*, *12*, 727-40.

70. Neuenschwander, B,, Branson, M., & Spiegelhalter, D.J. (2009) A note on the power prior. *Statistics in Medicine*, 28(28): 3562–3566. DOI: 10.1002/sim.3722.

71. O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... & Rakow, T. (2006). Uncertain judgements: eliciting experts' probabilities. John Wiley & Sons.

72. Pawlowicz, F., Gajewski, B.J., Coffland, V., Boyle, D., Bott, M., & Dunton, N. (2012). Application of Bayesian methodology within an equivalence content validity study. *Nursing Research*, *61*, 181-187.

73. Parreco, L. K., DeJoice, R. W., Massett, H. A., Padberg, R. M., & Thakkar, S. S. (2012). Power of an effective clinical conversation: Improving accrual onto clinical trials. *Journal of Oncology Practice*, 8(5), 282-286.

74. Philipson TJ, Mozaffari E, Maclean JR. Pharmacy cost sharing antiplatelet therapy utilization and health outcomes for patients with acute coronary syndrome. American Journal of Managed Care 2010; 16(4): 290-297.

75. Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March (pp. 20-22).

76. Polit, D.F., & Beck, C.T., (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*, 489-497.

77. Press, S. J. (2009). Subjective and objective Bayesian statistics: principles, models, and applications (Vol. 590). John Wiley & Sons.

78. R Development Core Team (2012). A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

79. Rashed, M. G., & Ahsan, R. (2012). Python in Computational Science: Applications and Possibilities. *International Journal of Computer Applications*, 46(20), 26-30.

80. Rosseel, Y. (2010). lavaan: Latent Variable Analysis. R package version 0.3-1. http://CRAN.R-project.org/package=lavaan.

81. Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Satistical Software*, *48*. 1-36

82. Samaniego, F. J., & Reneau, D. M. (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association,* 89, 947-957.

83. Schroen, A. T., Petroni, G. R., Wang, H., Gray, R., Wang, X. F., Cronin, W., ... & Slingluff, C. L. (2010). Preliminary evaluation of factors associated with premature trial closure and feasibility of accrual benchmarks in phase III oncology trials. Clinical Trials, 7(4), 312-321.

84. Speight, J., & Barendse, S.M., (2010). FDA guidance on patient reported outcomes. *British Medical Journal, 340,* c2921.

85. Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). Bayesian approaches to clinical trials and health-care evaluation. John Wiley & Sons.

86. Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2007). OpenBUGS user manual, version 3.0. 2. MRC Biostatistics Unit, Cambridge.

87. Stan Development Team. (2014). Stan Modeling Language Users Guide and Reference Manual, Version 2.2. http://mc-stan.org/

88. Stone, L. D., Keller, C. M., Kratzke, T. M., & Strumpfer, J. P. (2014). Search for the Wreckage of Air France Flight AF 447. Statistical Science, 29(1), 69-80.

89. Sullivan, S. G., & Greenland, S. (2013). Bayesian regression in SAS software. *International journal of epidemiology* 2013; 42: 308–17.

90. Team, R. C. (2011). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012. Open access available at: http://cran. r-project. org.

91. Thomas, J. H., & Nan, M. L(2009). fgui: A Method for Automatically Creating Graphical User Interfaces for Command-Line R Packages. *Journal of Statistical Software, 30*, 1-14. URL http://www.jstatsoft.org/v30/i02/.

92. van der Wouden, J. C., Blankenstein, A. H., Huibers, M. J., van der Windt, D. A., Stalman, W. A., & Verhagen, A. P. (2007). Survey among 78 studies showed that Lasagna's law holds in Dutch primary care research. *Journal of clinical epidemiology*, 60(8), 819-824.

93. Westland, J.C. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, *9*, 476–487.

94. Westland, J. C. (2012).Erratum: Erratum to Lower bounds on sample size in structural equation modeling [Electron. Commerce Res. Appl. 9 (6)(2010) 476-487]. *Electronic Commerce Research and Applications*, 11(4), 445.

95. Wilks, S.S. (1962). *Mathematical Statistics*. New York: Wiley.

96. Williams, H. C., & Seed, P. (1993). Inadequate size of 'negative'clinical trials in dermatology. *British Journal of Dermatology*, 128(3), 317-326.

97. Zhang, X., & Long, Q. (2012). Modeling and prediction of subject accrual and event times in clinical trials: a systematic review. *Clinical Trials*, 9(6), 681-688.

98. Zhang, X., & Long, Q. (2012). Joint monitoring and prediction of accrual and event times in clinical trials. *Biometrical Journal*, 54(6), 735-749.

99. Zhang, X., & Long, Q. (2012). Bayesian modeling and prediction of patient accrual in multi-regional clinical trials. Technical Report 12-01, Department of Biostatistics and Bioinformatics, Emory University.